A computer-assisted approach to supporting taxonomical classification of freshwater green microalga images

Vinícius Ruela Pereira Borges

Vinícius Ruela Pereira Borges

A computer-assisted approach to supporting taxonomical classification of freshwater green microalga images

Doctoral dissertation submitted to the Instituto de Ciências Matemáticas e de Computação – ICMC-USP, in partial fulfillment of the requirements for the degrees Doctorate Computer Science and Computational Mathematics. *FINAL VERSION*

Concentration Area: Computer Science and Computational Mathematics

Advisor: Profa. Dra. Maria Cristina Ferreira de Oliveira

USP – São Carlos
December 2016

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados fornecidos pelo(a) autor(a)

Borges, Vinícius Ruela Pereira
B634a A computer-assisted approach to supporting taxonomical classification of freshwater green microalga images / Vinícius Ruela Pereira Borges; orientadora Maria Cristina Ferreira de Oliveira. - São Carlos - SP, 2016.
181 p.

Tese (Doutorado - Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2016.

1. visual incremental classification. 2. green microalga images. 3. shape-based feature extraction. 4. image segmentation. I. Oliveira, Maria Cristina Ferreira de, orient. II. Título.

Vinícius Ruela Pereira Borges

Uma abordagem computacional para apoiar a classificação taxonômica de imagens de microalgas verdes de água doce

Tese apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como parte dos requisitos para obtenção dos títulos de Doutor Computer Science and Computational Mathematics. *VERSÃO REVISADA*

Área de Concentração: Computer Science and Computational Mathematics

Orientadora: Profa. Dra. Maria Cristina Ferreira de Oliveira

USP – São Carlos
Dezembro de 2016

I dedicate this work to my father, Ronaldo César Borges, and to mother, Janise Ruela Pereira Borges.

ACKNOWLEDGEMENTS

This is the end of another cycle in my life. I would like to thank God for everything, for giving me strength and health to be able to accomplish several missions in the last years. My family is also a part of that and I would like to thank them from the bottom of my heart for their support.

I would like to thank my advisor, Professor Maria Cristina Ferreira de Oliveira, for receiving me as her PhD. candidate, for the patience during the last years, for the instructions when we were conducting this work and for the support during my evolution as a researcher. I am very pleased to be able to attend to meetings, discuss research and share ideas. I also would like to thank Prof. Bernd Hamann for receiving me as a visiting research scholar in UC Davis and for taking your time with me to supervise my research during my time in United States.

I wish to acknowledge the finnancial support provided by FAPESP - the State of São Paulo Research Funding Agency, procs. 2012/00269-7 and 2013/26647-0.

I also would like to thank the support of my friends from Uberlândia, São Carlos, Davis (Califórnia) and Lavras. There have been so many years in this journey and friends are always a part of that. I'll never forget the good and bad times and I'll take you in my heart for the rest of my life.

“However difficult life may seem,
there is always something you can do and succeed at. ”
(Stephen Hawking)

ABSTRACT

BORGES, V. R. P.. **A computer-assisted approach to supporting taxonomical classification of freshwater green microalga images**. 2016. 181 f. Doctoral dissertation (Doctorate Candidate Computer Science and Computational Mathematics) – Instituto de Ciências Matemáticas e de Computação (ICMC/USP), São Carlos – SP.

The taxonomical identification of freshwater green microalgae is highly relevant problem in Phycology. In particular, the taxonomical identification of samples from the Selenastraceae family of algae is considered particularly problematic with many known inconsistencies. Biologists manually inspect and analyze microscope images of alga strains, and typically carry out several complex and time-consuming procedures that demand considerable expert knowledge. Such practical limitations motivated this investigation on the applicability of image processing, pattern recognition and visual data mining techniques to support the biologists in tasks of species identification. This thesis describes methodologies for the classification of green alga images, considering both traditional automated classification processes and also a user-assisted incremental classification process supported by Neighbor Joining tree visualizations. In this process, users can interact with the visualizations to introduce their knowledge into the classification process, e.g. by selecting suitable training sets and evaluate the results, thus steering the classification process. In order for visualization and classification to be feasible, accurate features must be obtained from the images capable of distinguishing between the different species of algae. As morphological shape properties are a fundamental property in identifying species, suitable segmentation and shape feature extraction strategies have been developed. This was particularly challenging, as different alga species share common morphological characteristics. Two segmentation methodologies are introduced, in which one relies on the level set method and the other is based on the region growing principle. Although the contour-based approach is capable of handling the uneven conditions of green alga images, its computation is time-consuming and not suitable for real time applications. A specialized formulation of the region-based methodology is proposed that considers the specific characteristics of the green alga images handled. This second formulation was shown to be more efficient than the level set approach and generates highly accurate segmentations. Once accurate alga segmentation is achieved, two descriptors are proposed that capture alga shape properties, and also an effective general shape descriptor that computes quantitative measures from two signatures associated to the shape properties. Experimental results are described that indicate that the proposed solutions can be useful to biologists conducting alga identification tasks once it reduces their effort and attains satisfactory discrimination among species.

Key-words: visual incremental classification, green microalga images, shape-based feature extraction, image segmentation.

RESUMO

BORGES, V. R. P.. **A computer-assisted approach to supporting taxonomical classification of freshwater green microalga images**. 2016. 181 f. Doctoral dissertation (Doctorate Candidate Computer Science and Computational Mathematics) – Instituto de Ciências Matemáticas e de Computação (ICMC/USP), São Carlos – SP.

A identificação taxonômica de algas verdes de água doce é um problema de extrema relevância na Ficologia. Identificar espécies de algas da família Selenastraceae é uma tarefa complexa devido às inconsistências existentes em sua taxonomia, reconhecida como problemática. Os biólogos analisam manualmente imagens de microscópio de cepas de algas e realizam diversos procedimentos demorados que necessitam de conhecimento sólido. Tais limitações motivaram o estudo da aplicabilidade de técnicas de processamento de imagens, reconhecimento de padrões e mineração visual de dados para apoiar os biólogos em tarefas de identificação de espécies de algas. Esta tese descreve metodologias computacionais para a classificação de imagens de algas verdes, nas abordagens tradicional e baseada em classificação visual incremental com participação do usuário. Nesta última, os usuários interagem com visualizações baseadas em árvores filogenéticas para utilizar seu conhecimento no processo de classificação, como por exemplo, na seleção de instâncias relevantes para o conjunto de treinamento de um classificador, como também na avaliação dos resultados. De forma a viabilizar o uso de classificadores e técnicas de visualização, vetores de características devem ser obtidos das imagens de algas verdes. Neste trabalho, utiliza-se extração de características de forma, uma vez que a taxonomia da família Selenastraceae considera primordialmente as características morfológicas na identificação das espécies. No entanto, a obtenção de características representativas requer que as algas sejam precisamente segmentadas das imagens. Esta é, de fato, uma tarefa altamente desafiadora considerando a baixa qualidade das imagens e a maneira pelas quais as algas se organizam nas imagens. Duas metodologias de segmentação foram introduzidas: uma baseada no método Level Set e outra baseada no algoritmo de crescimento de regiões. A primeira se mostrou robusta e consegue identificar com alta precisão as algas nas imagens, mas seu tempo de execução é alto. A outra apresenta maior precisão e é mais rápida, uma vez que as técnicas de pré-processamento são especializadas para as imagens de algas verdes. Uma vez segmentadas as algas, dois descritores para caracterizar as imagens foram propostos: um baseado em características geométricas básicas e outro que utiliza medidas quantitativas calculadas a partir das assinaturas de forma. Resultados experimentais indicaram que as soluções propostas têm um bom potencial para serem utilizadas em tarefas de identificação taxonômica de algas verdes, uma vez que reduz o esforço nos procedimentos manuais e obtém-se classificações satisfatórias.

Palavras-chave: classificação visual incremental, imagens de microalgas verdes, extração de características baseadas em forma, segmentação de imagens.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

ADF ..... Anisotropic diffusion filter
ANN ..... Artificial neural network
CMV ..... Coordinated Multiple Views
CSS ..... Curvature Scale Space
EM ..... Expectation-Maximization
FDM ..... Finite difference method
FN ..... False Negative
FP ..... False Positive
FPR ..... False Positive rate
GT ..... Ground-truth
GUI ..... Graphical user interface
HSV ..... Hue Saturation Value
IFT ..... Image Forest Transform
KNN ..... K-Nearest neighbors
MAT ..... Medial axis transform
MBR ..... Minimum bounding rectangle
MLP ..... Multilayer Perceptron
NJ ..... Neighbor Joining
PCA ..... Principal component anaylsis
PDE ..... Partial differential equation
RGB ..... Red Green Blue
SID ..... Segment Intersection Descriptor
SOM ..... Self-Organizing Maps
SVM ..... Support vector machine
TN ..... True Negative
TP ..... True Positive
TPR ..... True Positive rate
VCS ..... Visual classification system

LIST OF SYMBOLS

I — Digital image

I — Number of color channels

g — potential edge function

$\mathsf{d_i}$ — distance function, in which *i* is the identifier

μ — mean

Ω — image domain

V — Matrix of eigenvectors

$\mathsf{B_M}$ — mask

Γ — Parametric curve

Ω_1 — Foreground (regions of interest in images)

$\Omega \backslash \Omega_1$ — Background

ϕ — *Lipschitz* function

$\mathsf{d_E}$ — Euclidean distance

Σ — Covariance matrix

$\mathsf{B_i}$ — Binary image, in which *i* is the identifier

H — Heaviside function

$\mathsf{P_i}$ — Probability distribution of region *i*

θ_i — Distribution parameters of region *i*

$\mathsf{c_i}$ — Constant of region *i*

σ — Standard deviation

$\mathsf{s_i}$ — Seed pixel with respect to region identifier *i*

$\mathscr{C}$ — parametric representation of a closed contour

C — discrete representation of a contour

$\mathsf{n_p}$ — number of points in a discrete contour

κ — curvature function

$\mathscr{H}$ — Convex hull

CONTENTS

CHAPTER

1

INTRODUCTION

1.1 Context

Freshwater green microalgae play an important role in nature and human life. These microorganisms affect water properties such as color, odor and taste and interact with chemical compounds that can be potentially hazardous to human or animal health. Also a food source and primordial oxygen producers in aquatic environments [14], they are highly sensitive to environmental changes and therefore can signal the deterioration of ecological conditions, thus acting as effective indicators of water quality [15].

Researchers have been studying the potential of microalgae as biomass [16] or protein sources [17], in chemical processes [18], in oil production [19], and in medicine [20]. There is a huge variety of microalga species and families worldwide, and their taxonomical identification is a highly relevant problem in phycology. Recent taxonomical studies based on the 18S rDNA phylogeny [21] of freshwater green microalgae revealed an unknown diversity, especially in the Selenastraceae family, already known as possessing a highly problematic taxonomy [22] [21]. This has motivated biologists to concentrate additional efforts into deriving precise identification strategies for the Selenastraceae species, once this information is relevant for further investigations.

The traditional taxonomical identification of green microalgae is based on the analysis of morphological characteristics and it is often carried out manually by an expert. The procedure requires sampling alga cultures or environmental samples for observation under an optical microscope and then categorizing the observed organisms according to a predefined set of "identification keys" which essentially contemplate their morphological characteristics as their life cycle develops. This is a highly complex and time consuming process, demanding a detailed manual analysis of multiple images in order to identify the distinguishing characteristics of the various species. Furthermore, the accuracy of a suggested taxonomical identification is highly dependent on the taxonomist's training and expertise. Even for experts, the task may prove diffi-

cult and error prone, as some alga species share similar morphological characteristics, rendering their proposed classification inherently inaccurate.

Alternatively, an emerging molecular tool for species identification, the DNA barcode, has been successfully employed to distinguish among species and identify new species [23]. DNA barcoding has the advantage of being an objective tool for species identification in situations where identification is ambiguous and can be useful to delimit species that are morphologically similar [24]. Despite its potential, DNA barcode computation is not integrated into traditional taxonomical practices [22].

A major concern when identifying species of the Selenestraceae family is the subjective nature of the shape characteristics considered and the wide morphological variety of alga shapes. This leads to inconsistencies in the algae taxonomy, as some characteristics that identify particular species are also applicable to others [25]. Moreover, some researchers believe that current knowledge about the diversity of Selenastraceae is not well-explored on a world scale [22] [26] [27]. An ideal and promising approach to address this problem is to concentrate efforts on considering several sources of alga characteristics, such as multiple genes that are more variable than the 18S rDNA. For that purpose, researchers worldwide should standardize the identification criteria and work on building a concrete taxonomical basis using the traditional taxonomy, phylogenetic relations, biochemical data, digital images and observing the behavior and evolution of cultured alga cells.

1.2 Motivation

There is much interest in investigating computational techniques for supporting taxonomical classification tasks [28] [29] [6]. In particular, image processing, feature extraction, pattern recognition and visual data mining provide potential approaches to address the taxonomical identification of green alga species. Image processing is potentially useful to improve the image conditions and segment target algae from digital microscope images [30] [31]. Feature extraction techniques provide methods to extract and describe relevant physical properties of alga regions and to represent them in a suitable structure for further processing [32], whereas pattern recognition supports classification, clustering and learning approaches useful for species identification [33]. Visual data mining aims to combine the human capability of recognizing visual patterns with automatic or semi-automatic mining techniques in order to maximize human performance on data analysis tasks [34] [35].

There have been previous efforts towards developing computational support for taxonomical identification of alga species, such as diatoms, phytoplankton and other microrganisms [9] [36] [6]. Typically, the systems embed image processing and pattern recognition algorithms that capture the relevant image properties and derive an appropriate representation for further processing. However, those methodologies were developed aiming to classify particular alga species, so the morphological characteristics of the organisms and their taxonomy are

specific to the species considered. Using these methodologies for feature extraction and identification of the Selenastraceae algae would result in low identification accuracy as the features would not be sufficiently representative. Furthermore, the target user (a biologist) would be required to adjust classifier settings, build training sets, interpret classification results. However, such tasks may be unfamiliar to them.

Previous studies have shown that automatic classification in general is highly dependent on the nature of data and on the learning approach [2] [37] [38]. Generally, data may present high dimensionality and the similarity relationships between data instances are not taken into account when setting some classification tasks, such as building the training set [38]. Hence, by adopting a visual exploration process supported by interactive tools, users can participate in the classification either by analysing the visualizations or introducing their knowledge into the process. For instance, one could select suitable examples for the training set and adjusting classifier parameters [2] [35].

The current state-of-art in pattern recognition from images and recent results in visual analytics motivate further investigations on how such strategies can be applied to facilitate the task of biologists conducting taxonomical identifications. Particularly, visualizations based on phylogenetic trees have been proved useful to support users in interpreting and understanding the relationships amongst images from a collection [39]. Such tree-based visualizations have also been successfully applied in incremental image classification [2] [40], which includes the users into classification loop by allowing them to handle visualizations and use their knowledge to build training sets or adjust classifier settings. This was an motivation considering a similar strategy to the problem of green microalgae classification.

However, this requires transforming the alga images into a representation that can be used as input to the visualization and classification methods. For this purpose, this work proposes two feature descriptors that characterize the visual patterns of green algae as they appear in the digital images. The focus is on shape-based features, since shape is the most proeminent visual property considered by biologists in discriminating green alga species [41]. The descriptors are formed by basic geometric features and additional measurements derived from other complex shape descriptors. Experimentation has shown the effectiveness of these descriptors in distinguishing green algae and other kinds of shapes.

The automatic feature extraction requires segmenting the alga shapes. Segmentation of alga cells must be highly precise, so that no fine detail relevant to distinguish between species in further classification stages is missed out. Thus, this research also describes strategies for automatically and accurately segmenting alga regions from microscope images. Two segmentation methodologies are proposed: one based on the level set method [42] and the other based on the region growing principle [43]. The latter presents better segmentation accuracy since it is specialized in processing the Selenastraceae alga images.

1.3 Goals and Contributions

The main goal of this research is to explore the applicability of image processing, pattern recognition and visual data mining approaches in tasks of taxonomical identification of green microalgae images. Achieving this goal required developing techniques for image segmentation, automatic feature extraction and approaches for an accurate classification of alga species of the Selenastraceae family to reduce human effort and improve correctness while retaining user control.

Other specific objectives are accomplished, listed below:

1. Define a user-guided incremental classification process with the support of tree-based similarity visualizations of green microalga images. This strategy contemplates the presence of a biologist into the computer-assisted classification process, as well as the use of traditional classifiers.

2. Develop automated and highly specialized shape feature extractors, so that green algae can be described according to their shape patterns and morphological characteristics. As a result, a green algae feature descriptor formed by some selected shape-based measures is obtained for use in further classification and visualization tasks.

3. The development of a segmentation methodology to produce representative green alga shapes from the microscope images. The methodology should preserve to the highest possible accuracy the diverse alga shape patterns.

The main contributions of this thesis are:

1. Two automatic segmentation methodologies for green alga images and similar biological images based on: a propagation of a closed and dynamic curve; a region growing principle.

2. A green microalgae feature descriptor capable of precisely characterizing morphological features and cell organizations of algae from the Selenastraceae family.

3. A simple shape descriptor of global nature and low computational cost that computes statistical measures from two signatures of a given shape.

4. The study and application of two strategies for classifying species of green microalgae based on: traditional classifiers on conventional classification processes; user-guided incremental classification with the support of tree-based visualizations.

This research involved the application of techniques from image processing, pattern recognition and visual data mining to a real and difficult problem in phycology. This required

extensive tuning of previous general-purpose solutions to make them effective for automatically segmenting, describing and classifying species of algae from the Selenastraceae family. The resulting segmentation approaches have been devised specifically for this problem, but may be applicable to other similar scenarios that may require precise cell shape segmentation from images acquired from optical microscopes. The shape extractors were also shown to be effective in benchmark shape image sets.

From the point of view of the biologists, this research contributed with a computer-assisted approach to supporting the taxonomical identification of green algae from the Selenastraceae family based on the analysis of digital microscope images. The proposed solutions rely on a morphological analysis of the alga characteristics, as observed in the images, relating them with specific shape features automatically extracted. The classification approach can help biologists to make decisions in situations that species identification is ambiguous, or when there is divergence between the traditional morphological analysis and molecular phylogeny, besides reducing the manual effort required whilst retaining their control over the process.

1.4 Organization

This thesis is organized as follows: Chapter 2 contextualizes the current scenario of taxonomical identification of freshwater green microalgae, provides the fundamentals of image processing, pattern recognition and visual data mining, as well as a survey of related work. Chapter 3 presents the proposed methods for accurate segmentation of green algae images: one based on the level set method and the other based on a region growing principle. Chapter 4 describes the proposed shape-based feature extractors, which led to the definition of two shape descriptors for green algae. Some experiments are also reported to evaluate their discrimination capability in relation to the morphogical characteristics of the several green alga species considered. Chapter 5 describes the approaches for automatic classification of green microalga images, as well as a user-guided incremental classification with the support of tree-based visualizations. Moreover, it describes experiments conducted to evaluate the performance of the automatic classification and a case study involving a biologist conducting a visual incremental classification. Finally, Chapter 6 provides conclusions, limitations and discusses possible future research.

CHAPTER

2

BACKGROUND AND RELATED WORK

The taxonomical identification of green microalga species is complex and time consuming due to the detailed analysis of algae morphological characteristics, and the inconsistent taxonomy of the Selenastraceae family. Recent developments related to image processing, pattern recognition and visual data mining can support the biologists in this process by detecting alga cells in digital images, computing features describing their morphological characteristics, and performing classification at the species level.

This chapter aims to contextualize the scenario faced by the biologists when identifying freshwater green microalgae of the Selenastraceae family, and present potential computer vision methods to address the problem. Section 2.1 describes some concepts of taxonomy and the current practices for taxonomical identification of green alga species. Moreover, the morphological characteristics of the green alga species covered in this research are described. Section 2.2 introduces some fundamentals of image processing, feature extraction, classification and visual data mining. Section 2.3 presents related research on literature concerning the classification of organisms in biological or microscope images, as well as visual exploration processes applied to similar problems in biology.

2.1 Freshwater green microalgae

Organisms are grouped together into taxa (singular: taxon) and given a taxonomic rank, groups can be aggregated to form a super group of higher rank and thus create a taxonomic hierarchy. The current taxonomical rank comprises 7 ranks, making the sequence kingdom, phylum (or division), class, order, family, genus and species, in which *Kingdom* is the second highest rank and Species is the most restrict one. Taxonomy is the science that defines and labels groups of biological organisms based on their shared characteristics.

Phycology is a science field and branch of botany that studies algae life. Algae is a term used to represent a large and diverse group of eukaryotic organisms [44]. Many species

are unicellular and microscopic (including *Chlorella*, phytoplankton and other microalgae), as shown by the diatom cell depicted in Figure 1(a). Others alga organisms are multicellular to one degree or another, some of these growing to large sizes, such as the giant kelp depicted in Figure 1(b), a large brown alga, which is distinguished from the higher plants by the lack of true roots, stems or leaves.

(a) (b) (c)

Figure 1 – The wide diversity of alga organisms: (a) microscopic image of diatoms [1]; (b) a giant kelp [2]; (c) indication of the amyloid pyrenoid in a microalga.

However, this definition of algae in general is rather based on their appearance and functioning, not taking into account that they are a highly diverse and heterogeneous group of organisms with varying phylogenetic origins [45]. In particular, microalgae can be characterized by their small size as compared to other groups and the ability to perform photosynthesis, thus including photosynthetic eukaryotes [46]. These microorganisms can be taken as indicators to monitor freshwater ecosystem condition because they react quickly and predictably to a broad range of pollutants and environmental changes [47] [15].

Algae are an important group of living organisms from an ecological perspective, once they form the base of the food chain in oceans, rivers and lakes [48]. Along the last years, freshwater green microalgae have been exploited as promising candidates for environmental studies, for understanding organisms lives and in biotechnological applications considering their use as feedstock for biofuel production [49], among others [16] [18] [19] [20]. These studies usually consider particular microalga species due to their benefits for a target application, and since specific species are suitable for experimentation and case studies. Thus, the correct identification of a particular species is important to evaluate the phylogenetic relationship with other algal groups, being an indispensable information for taxonomical and ecological studies.

2.1.1 Taxonomical identification

There is a huge variety of microalga species and families worldwide, and their taxonomical identification is a highly relevant problem in phycology. Professor Armando Augusto Henriques Vieira, from the "Departamento de Botânica" at Federal University of São Carlos (UFSCar) supervises a research project under the "Programa Biota" funded by FAPESP (São Paulo Research Foundation), which addresses the study of the biodiversity of freshwater microalgae

[1] http://people.westminstercollege.edu/faculty/tharrison/emigration/2_diatoms.gif
[2] http://www.nature.com/nature/journal/v461/n7267/images/4611066a-i1.0.jpg

(FAPESP Process 2011/50054 – 4) [3]). The purpose of this project is to create a germplasm bank, cryopreserved in liquid nitrogen, of freshwater microalgae (phytoplanktonic, but also thycoplanktonic, benthic and sub-aerial which may be cultivated and frozen), in order to preserve the biodiversity of these organisms, along with a database of information on their species.

Recent taxonomical studies of freshwater green microalgae have revealed the occurrence of a yet unknown diversity, particularly the algae of the Selenastraceae family, known as possessing an inconsistent taxonomy, especially in species of genera *Ankistrodesmus*, *Monoraphidium*, *Selenastrum* and *Kirchneriella* [22] [21]. On the other hand, standard taxonomical identification that relies on the analysis of morphological characteristics of specimens collected in the environment is highly problematic, due to the subjective nature of the features considered and the morphological variety of alga shapes, which may be only revealed in cultured strains [50] [51] [52]. For instance, the presence of amyloid pyrenoid, a feature considered for differentiation at the species level (Figure 1(c)), sometimes can only be observed through electron microscopy, or in some cases it may appear or not in a particular species depending on the culture conditions. Phylogenetic studies indicated that morphologically similar strains may be molecularly very distinct, and also the opposite, with molecularly similar strains showing diverse morphology [21] [53]. Experts recognize that current knowledge on the specific diversity and ecology of the Selenastraceae in a worldwide scale is very limited. It is also fundamental to improve the effectiveness of standard approaches that rely on analyzing morphological properties.

The discrimination of alga species by taxonomists studying the evolution of such organisms is a complex task, requiring simultaneous analysis of data from morphology, ecology and behavior in the environment, as well as their DNA [21]. Particularly, the DNA barcoding method [54] is a molecular barcode used for recognizing known species, reducing the usage of ecological or morphological data, which is less reliable. The method has standardized steps concerning the organisms, which allows identifying organisms at several life stages. The ultimate goal is to build a unified database containing a set of all possible species [4].

However, only the DNA barcode is not sufficient to differentiate highly correlated species because its computation is currently not integrated into traditional taxonomical practices [22]. The species which have the DNA barcode extracted should be inserted in a trustable taxonomical system, because the differentiation in some groups cannot be obtained without errors. This is the case of freshwater green microalgae, which present high morphological convergence, polymorphism and a high incidence of critical species. In this scenario, Prof. Vieira's group has been conducting studies under the Biota program that combine traditional taxonomy and the sequencing analysis of the 18S rDNA genes to identifying freshwater green microalgae of the Selesnastraceae family. Their goal is to provide a taxonomical basis to identify the biodiversity of those organisms, which is highly problematic according to taxonomists [41].

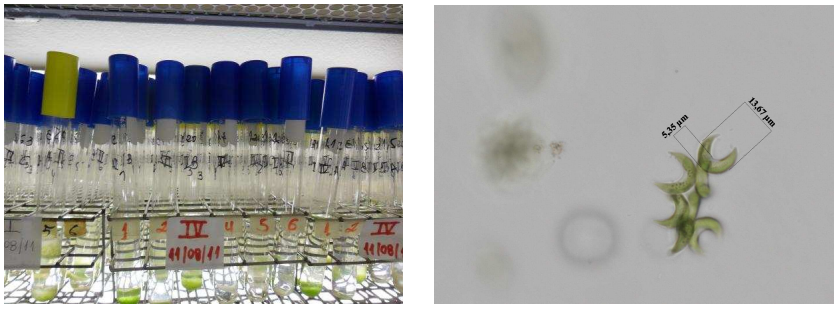
[3] Title (in Portuguese): Biodiversidade de microalgas de Água doce: banco de germoplasma e obtenção de marcadores moleculares das espécies criopreservadas

[4] http://www.barcodinglife.org/

Thus, the DNA barcode associated with practices from traditional taxonomy can lead to a more accurate identification and recognition of microalga species. In some cases, the morphological features are extremely helpful as a starting point and, in many cases, they are fundamental for species discrimination.

2.1.2 Morphological analysis

Prof. Vieira's research group owns a private collection of alga cultures [5], as illustrated in Figure 2(a). Each test tube stores an algae culture of a unique species and its label records the starting date of culture and the associated species. Algae are cultivated in suitable conditions for their growth, regarding temperature, photoperiod and culture medium. The biologists regularly sample each culture to inquire the state of the cultured strain and check for contamination. Afterwards, alga cells are viewed under a light microscope with an attached digital camera. In the microscope, the biologists analyze some properties of alga cells, such as depth, width, thickness, colony formation, cell texture, color and the density of mucilage, and then digitalize the current view to generate an image file. The morphological characteristics of green algae are computed using the software *Axiovision* [6] from the acquired image. Once the morphological characteristics are computed, identification is performed according to the traditional taxonomy of green algae from the Selenastraceae family as given by the identification keys (further described). This process occurs several times during the algae life cycle, as the morphological characteristics of algae can vary during their life cycle.



(a) (b)

Figure 2 – Conventional taxonomical identification: (a) several algae culture maintained in laboratory; (b) process of computing measures of alga cells in a digital image using the *software Axiovision*.

The taxonomical identification of green microalgae at the species level considers other image details that can also be explored with *Axiovision*, such as cell width and height, shown in Figure 2(b). Furthermore, the algae shape (ellyptical, fusiform, elongated, round, and other subtle variations), the shape apexes (round or sharp), the occurrence of colonies, cell division, the presence or absence of mucilage, and the presence of amyloid pyrenoid are observed during the visual analysis. This whole set of morphological characteristics are compared with predefined

[5] CCMA-UFSCar (Coleção de Culturas de Microalgas de Água doce da Universidade Federal de São Carlos).
[6] AxioVision 4.6 (Carl Zeiss Group, Oberkochen, Germany)

morphological parameters, known as identification keys [25] [55] [56].

Identification keys are systematic descriptions of the morphology of biological entities, such as plants, animals, fossils, microorganisms, and pollen grains [57]. The process of identifying an alga species using an identification key resembles a binary decision tree [58]. Considering the Selenastraceae taxonomy, given a set of morphological characteristics and observations, the analysis proceeds along a path towards terminal branches that indicate the species. During this process, decisions are taken based on the morphological characteristics of the set of input features, where is split at multiple stages into branch-like segments based on a specific criteria.

Taxonomical identification based on morphology demands a detailed manual analysis of multiple images in order to identify the distinguishing characteristics of the various species. Even for experts, this task may prove difficult and error prone, as some alga species share similar morphological characteristics. Thus, computational approaches to automatically detecting, computing and selecting the relevant features of green algae captured in digital microscope images can support the biologists in such tasks.

The design and modeling of image processing and pattern recognition techniques for this purpose requires some previous knowledge about the green microalgae. Therefore, the next topics describe the species of the Selenastraceae family considered in this work and details their morphological characteristics that are important for extracting the key features for accurate taxonomical classification.

2.1.3 Selenastraceae family

The morphology of species recognized as belonging to the Selenastraceae family comprises a variety of cell shapes: from coccoid to elongated, cylindrical to fusiform, sickle-shaped to spirally curved, with sharp or rounded ends, where cell arrangements vary from solitary to colonial forms [25] [59]. Based on these criteria, up to 100 species have been described in various genera and included in this family [60] [21] [53]. Since 1903, the family's description has undergone many taxonomical changes.

The Selenastraceae taxonomy still needs revision, since morphological characteristics are usually not in accordance with molecular data [22] [53]. For instance, *Selenastrum bibraianum* (a Selenastrum specie), and *Selenastrum gracile* belong to different phylogenetic lineages according to 18S rDNA phylogeny [21] but no taxonomical changes were made in the genus, since the authors suggested further studies to verify these findings.

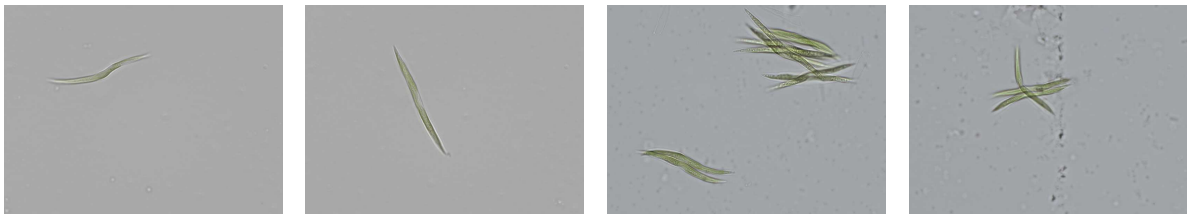
The following descriptions detail the species of the Selenastraceae family considered in this research.

Ankistrodesmus densus Korshikov, 1953

Ankistrodesmus densus is characterized by cells in colonies and sometimes solitary, and their size varies depending on environmental conditions [61]. The cells are cylindrical, suddenly

pointed, nearly straight and elongated [59]. Figures 3(a) and 3(b) illustrate two examples of solitary cells, one curved and other almost straight, respectively.

Colonies of *Ankistrodesmus densus* present either curved cells or cells twisted around each other in their central part, or overlapping. These groupings contain numerous cells, sometimes even several generations of cells. The younger cells usually form colonies, containing four cells, with a central twist. Figure 3(c) shows a colony with complex shape geometry and Figure 3(d) depicts a group of three algae joined by their central parts.



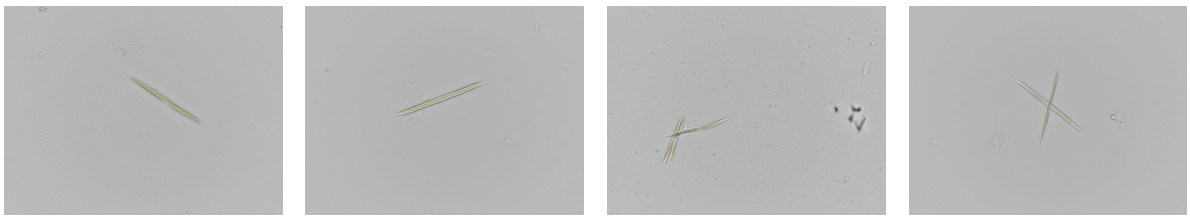
(a) (b) (c) (d)

Figure 3 – Samples of *Ankistrodesmus densus*: (a-b) single alga forms; (c-d) two distinct colony formations.

Ankistrodesmus fusiformis Corda ex Korshikov, 1953

According to Fott et al. [61], *Ankistrodesmus fusiformis* cells present fusiform or cylindrical cells, straight or nearly straight and are slightly curved to arcuate. Figures 4(a) and 4(b) show two cases of single alga.

Ankistrodesmus fusiformis colonies present an unstable cruciform or stellate geometric shape, with cells joined on their midregion only for a short period when adult, separating later into single cells. Figure 4(c) depicts an uncommon case of alga colony and Figure 4(d) presents these scenarios in which cells group and are joined in their centers.



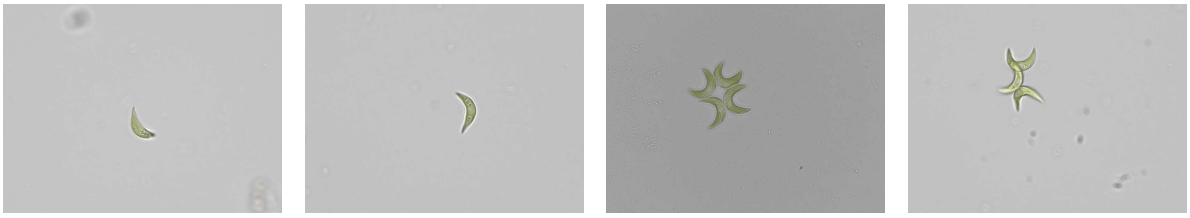
(a) (b) (c) (d)

Figure 4 – Samples of *Ankistrodesmus fusiformis*: (a-b) two examples of *Ankistrodesmus fusiformis* solitary cells; (c-d) two colony forms.

Selenastrum bibraianum Reinsh, 1866

Described as type-species for Selenastraceae, *Selenastrum bibraianum* presents C-shaped to fusiform cells. A single cell can have a low curvature geometry, having slightly rounded apexes, and it is gradually pointed, slightly rounded in autospores when in advanced age. Two solitary algae are depicted on Figures 5(a) and 5(b).

Colonies of *Selenastrum bribaianum* are cruciform in section and there are often two fascicles, one above the other, placed symmetrically along the central axis. These situations are depicted in Figures 5(c) and 5(d), respectively. The cells are clustered in fascicled colonies, joined in fours by their convex walls. The colonies are embedded in a strong layer of mucus and are of an almost globular shape.

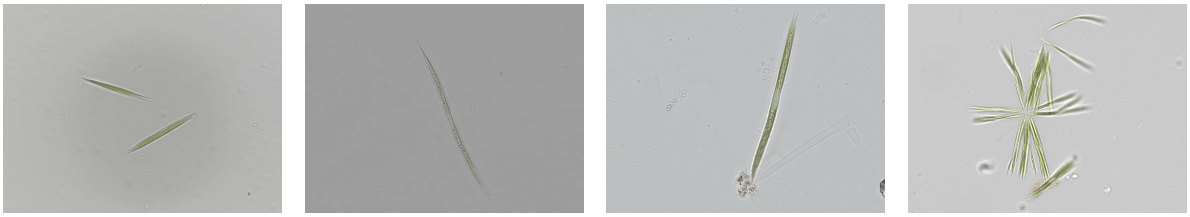


(a) (b) (c) (d)

Figure 5 – Samples of *Selenastrum bribaianum*: (a-b) examples of single cells; (c-d) two kinds of colony formation.

Monoraphidium griffithii (Berkeley) Komárkova-Legnerova, 1969

Monoraphidium griffithii cells have an interesting feature of varying their length according to the environmental conditions. The strain cells change by as much as one quarter of the average length [61]. This is a single cell species, with fusiform, cylindrical, longer than broad, straight, curved or sigmoid and sometimes spiral. Figures 6(a), 6(b) and 6(c) present, respectively, cells of small length, a larger cell without mucilage and a larger cell with the presence of mucilage in one of the corners. Sometimes cells group for reproducing by means of autospory, as shown in Figure 6(d). This species does not form colonies.



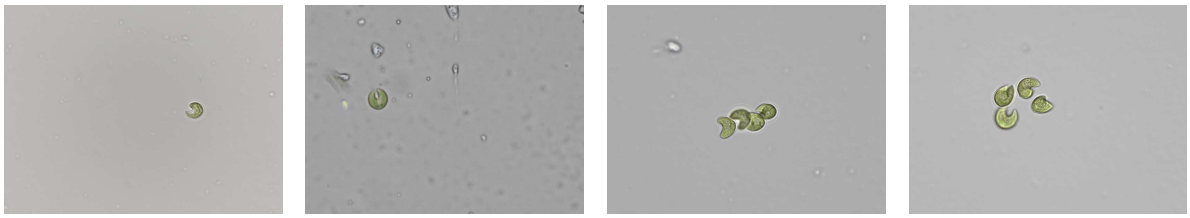
(a) (b) (c) (d)

Figure 6 – Samples of *Monoraphidium griffithii*: (a-b-c) single cell form; (d) cells reproducing by autospory.

Kirchneriella aperta Teiling, 1912

Kirchneriella aperta cells have a nearly circular outline with a usually broad V-shaped concavity, only occasionally with near parallel sides and rounded apexes or bluntly pointed. Thus, alga cells are strongly curved and crescent-shaped, rather thick. Figures 7(a) and 7(b) show examples of single cells of *Kirchneriella aperta*, in which the V-shaped cells with concavity can be noted.

Colonies are spherical to ellipsoid, composed of 4 to 16 cells irregularly arranged within the gelatinous envelope. Figures 7(c) and 7(d) depict examples of varied colony shapes.



(a) (b) (c) (d)

Figure 7 – Samples of *Kirchneriella aperta*: (a-b) single forms; (c-d) examples of colonies.

Raphidocelis subcapitata (Korshikov) (Korshikov) Nygaard, Komárek, J.Kristiansen & O.M.Skulberg, 1987

These microalgae are commonly used as a bioindicator species to assess the levels of nutrients or toxic substances in freshwater environments [62]. They are quite sensitive to the presence of metals and have a ubiquitous distribution, being thus widely employed in ecotoxicology [63].

Raphidocelis subcapitata is mainly present in a single form. Their shapes are curved and twisted, similar to a sickle. Figures 8(a) and 8(b) present these cases, while Figures 8(c) and 8(d) illustrate some colonies.

(a) (b) (c) (d)

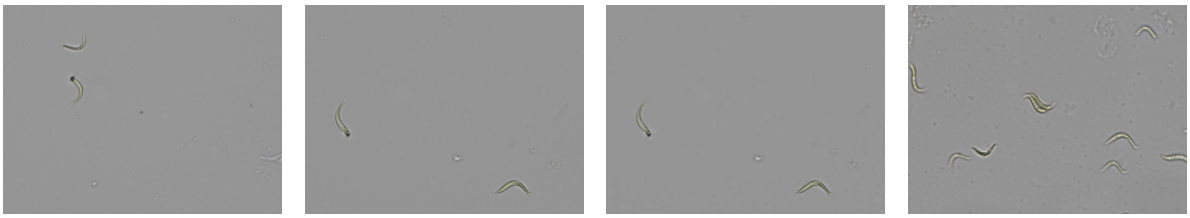
Figure 8 – Samples of *Raphidocelis subcapitata*: (a-b) single form examples; (c-d) assort of single and colony forms.

Monoraphidium contortum (Thuret) Komárková-Legnerová, 1969

The ecology of *Monoraphidium contortum* is varied, as they can be found in clean or polluted waters, and also in alkaline environments. Figure 9 illustrates some examples. Generally, these algae are found as isolated organisms, and their shapes are elongated, fusiform and gradually pointed. Figures 9(a), 9(b) and 9(c) depict their overall shape geometry, which is irregularly curved and similar to a sigmoid. Figure 9(d) shows two cells after assexual reproduction, known as autospory, a typical characteristic of the family [25].

2.2 Image processing and analysis

Let $\Omega \subset \mathbb{R}^2$ be the image domain and $I : \Omega \to \mathbb{R}^l$, where $l \in \mathbb{N}$. I is a function designed to describe a digital image, in which $I(x,y)$ records the light intensity at domain point (x,y).



(a) (b) (c) (d)

Figure 9 – Samples of *Monoraphidium contortum*: (a-b-c) three solitary cells; (d) cells in autospory.

Digital image processing techniques require the continuous representation I to be transformed into a discrete representation that can be handled by computers.

A digital image I is obtained from an analog image by means of sampling and quantization processes, called digitalization [64]. Sampling consists of discretizing the image domain in the vertical and horizontal directions, resulting in a 2D-array $m_1 \times m_2$. Each element $I(i, j), 0 \leq i \leq m_1$ and $0 \leq j \leq m_2$ is called a pixel. Assuming I is a monochromatic image ($l = 1$), the quantization process determines a number of gray levels that describe the image pixels. Generally, monochromatic (grayscale) images are described using 256 gray levels, in which brighter intensities are assigned to greater intensity values (closer to 255) and darker intensities to values closer to zero.

While in a grayscale image pixels are described by a single intensity value, in a color image (or RGB image), the pixel color is given by a vector, in which $l \geq 3$. The vector stores the intensities of the primary colors red (R), green (G) and blue (B), and the pixel color is obtained from their combinations.

In computer vision systems, the physical properties and visual characteristics of digital images are described by numerical attributes represented as feature vectors. Generally, those systems employ data mining and machine learning algorithms that require as input a simplified representation of data instead of using the digital image itself. The steps that transform a raw image prior to computing the numerical attributes of the feature vectors are further discussed.

2.2.1 Preprocessing: image filtering

Raw biological images are typically noisy, presenting low contrast, non-uniform illumination conditions, scale problems and artifacts. These are due to environmental and physical conditions of the image acquisition process. The direct application of feature extraction techniques on raw images can degrade the result of further processing steps, as their relevant characteristics may be concealed.

Preprocessing a raw image refers to manipulating the image to obtain a more appropriate result for a specific application or further processing [65]. Most techniques reported in the literature regarding biological images detail preprocessing and postsegmentation procedures to improve image quality and conditions. Some popular approaches rely on image denoising (or

noise removal) [66] [67] [68] [69], image (and edge) enhancement [70] [71] and noise suppressing [72].

Several image processing systems employ Gaussian or median filtering for noise suppression in preprocessing steps [64] [73]. Despite their simple formulation and low computational cost, they are not indicated in some cases since Gaussian filtering blurs region boundaries and the median filtering is more appropriate for applying to images with salt-and-pepper noise [64].

From a wide diversity of possible image filtering techniques, those based on anisotropic diffusion [74] have been successfully applied to images from several domains for noise supression, such as medicine [75] [76], Synthetic aperture radar (SAR) [77] and botany [78]. Perona and Malik [67] formulated a non-linear diffusion method for avoiding the blurring and localization problems of the aforementioned filtering techniques. Basically, the idea is to apply an inhomogeneous process that reduces the diffusivity at specific locations which is likely to be edge regions.

Perona-Malik's formulation has been receiving updates along the past years. In this work, the anisotropic diffusion filter (ADF) introduced by Barcelos et al. [76] has been applied to the original green alga images. The mathematical formulation of the ADF filter relies on a Partial Differential Equation (PDE), given by:

$$u_t = g|\nabla u| div\left(\frac{\nabla u}{|\nabla u|}\right) - \lambda(1-g)(u-I), \tag{2.1}$$

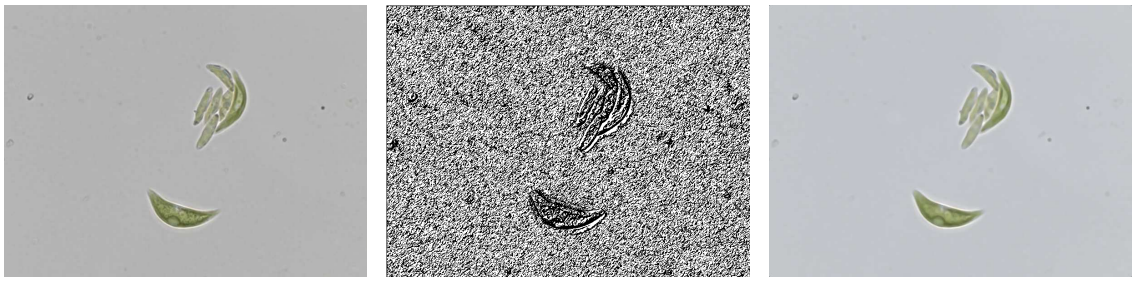
in which div is the divergent operator, $I = I(\mathbf{x})$ is the image to be filtered and $u = u(\mathbf{x},t)$ is a smooth version of I at a time step $t > 0$, where $u(\mathbf{x},0) = I(\mathbf{x})$ and $\mathbf{x} \in \Omega$. The parameter λ balances the smoothness of the region boundaries – the goal is to smooth the boundary whilst preserving the important shape properties. g is a positive boundary potential, usually chosen as a decreasing function of the image gradient. This function must satisfy $\lim_{r\to\infty} g(r) = 1$, so that the diffusion process is reduced in the boundaries. Thus, a usual choice for g is given by:

$$g(r) = \frac{1}{1+|\nabla r|^2}. \tag{2.2}$$

Eq. (2.1) is solved numerically by computing the Euler-Lagrange equations associated with a gradient descent scheme [79]. The partial differential equations obtained are discretized using the Finite Difference Method [80]. In order to compute the numerical solution for u, the parameters λ and number of iterations in Eqs. (2.1) and (2.2) are set according to the image characteristics and the desired degree of smoothness. Further details about ADF discretization and implementation and about parameter settings can be found elsewhere [74] [81].

Figure 10 illustrates this filtering process applied to a green algae image, shown in Figure 10(a). The image representing function g is depicted in Figure 10(b), in which boundaries

are related to darker pixels. The image smoothing result is presented in Figure 10(c), in which algae and background regions are more homogeneous and their edge information is preserved.



(a) (b) (c)

Figure 10 – Anisotropic diffusion filtering: (a) original RGB image; (b) function g represented as an image; (c) smooth image obtained by anisotropic diffusion.

The major drawbacks of the anisotropic diffusion filtering are the parameter settings and the computation of the numerical solution, which can be time consuming when performing several iterations.

2.2.2 Image segmentation

The goal of segmentation is to subdivide the image domain in its constituent regions, aiming to simplify its representation for further analysis [64]. Automatic segmentation is a highly complex task in image processing and feature extraction techniques depends on the segmentation quality to compute representative features. According to Gonzalez et al. [64], segmentation algorithms generally fall under two basic categories in handling intensity values: edge and region.

Edge-based methods assume that region boundaries are sufficiently different from each other and from background to allow boundary detection based on local intensity discontinuities. Algorithms in this category attempt to detect abrupt changes in image intensities, such as isolated points, lines or boundaries. Thus, edge-based approaches are not recommended on noisy images because these unwanted patterns are similar to discontinuities and can also degrade region boundaries. Popular approaches for detecting discontinuities are the Canny filter [82], the Hough Transform [83] [84] and the Laplacian of Gaussian (LoG) [85]. Other sophisticated techniques implement dynamic curves that propagate through the image domain until reaching region boundaries: active contours [86] and level sets [42].

Region-based segmentation methods attempt to group pixels that satisfy a homogeneity criterion of image regions [87]. Examples of homogeneity criteria include the texture of regions of interest or a specific range of gray level intensities. This approach is limited when images present low contrast or illumination problems, once it is highly dependent on intensity homogeneities. Region growing [43] [88], QuadTree [89] [90] and Watershed transform [91] [92][93] are examples of this category.

Most segmentation techniques for biological microscopy images are based on threshold,

edge and region approaches [94] [95] [96]. Figure 11 shows the segmentation results of the two images in Figures 11(a) and 11(e), in which standard approaches from the literature are applied to green microalga species of Selenastraceae. Figures 11(b) and Figure 11(f) show the result of a binarization on the green channel of the original images using a threshold value computed with Otsu's method [97]. The poor segmentation results are due to the wide range of intensities found in the interior of alga regions. Figures 11(c) and 11(g) illustrate the results with the Canny Edge detector to the green channel of the images in Figures 11(a) and 11(e), respectively. Noticeably, the contours detected are not closed. The abrupt changes in intensities in the alga regions prevent the method from obtaining closed and regular contours. Figures 11(d) and 11(h) are the segmentations obtained with the Watershed Transform, in which the area corresponding to the alga region in the watershed image are manually selected. This result is affected by the smooth intensity variation in the background and the transparencies in the alga corners, which lead to poor segmentation because some alga areas could not be correctly recognized.

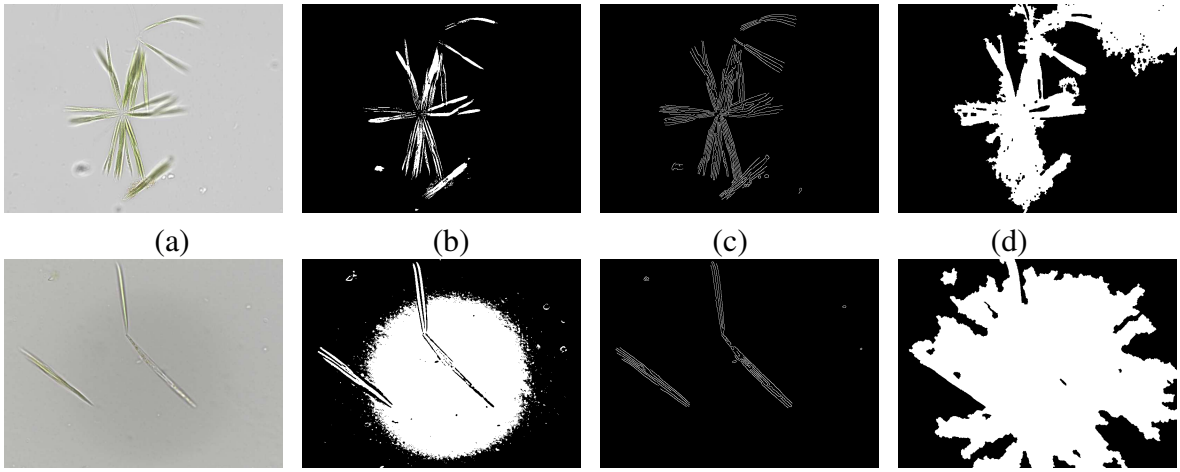


Figure 11 – Conventional approaches for segmenting biological images applied to two green alga images (one per row): (a-e) original green alga image; (b-f) result obtained with binarization using Otsu's threshold; (c-g) result obtained with edge detection using the Canny algorithm; (d-h) result obtained with the Watershed Transform.

These segmentation results show that such approaches cannot perform well in some cases due to the nature and characteristics of the alga images. Thus, their segmentation might require specific methodologies to deal with particularities, such as transparencies, blurred areas in alga cell body, noise and intensities that are not green. This work proposes two segmentation methodologies for Selenastraceae microalga images, described in Chapter 3: one based on the level set method and the other based on the region growing principle.

2.2.3 Automatic feature extraction

Feature extraction refers to the computation of quantitative measures that describe the visual patterns of an image for further steps, such as data mining, pattern recognition and analysis [98]. In general, it takes an input image (raw or segmented) and produces feature vectors

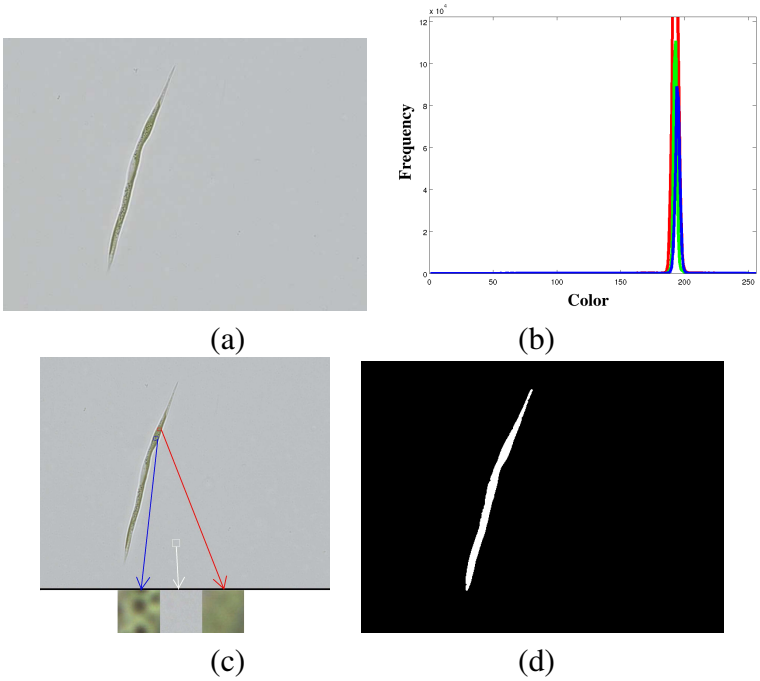


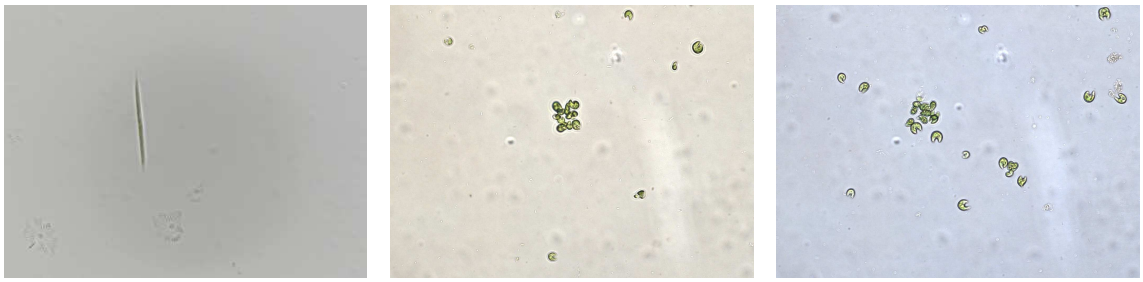
Figure 12 – Green microalga image: (a) Raw image; (b) Color histogram; (c) Different texture samples; (d) Alga shape.

as output. A set of those structures defines a feature space, in which the number of dimensions is equivalent to the number of computed features (or attributes). A "good" feature extractor should compute representative numerical attributes, i.e., attributes that are similar for images of the same class and distinct for images from different categories [99].

A digital image, or its regions obtained from a segmentation process, can be represented by means of its external characteristics (boundaries) or by its internal characteristics (pixels constituting the regions of interest) [64]. In the former, the image description considers region shapes and their internal objects. In the latter, visual properties of the image pixels are taken into account, such as color and texture.

Color is a property related to image regions and objects that describes the amount of reflected light in a scene. Light that is void of color is called monochromatic and intensity is the attribute that describes such property. On the other hand, chromatic (color) light is represented by three quantities: radiance, luminance and brightness [64]. Popular approaches for color description are the color histograms, such as the *Border/Interior Pixel Classification* (BIC) [100] and the *Color Coherence Vector* (CHV) [101] [102]. Figure 12(b) shows the intensities histogram for the image in Figure 12(a), which describes the intensities frequency of each color channel.

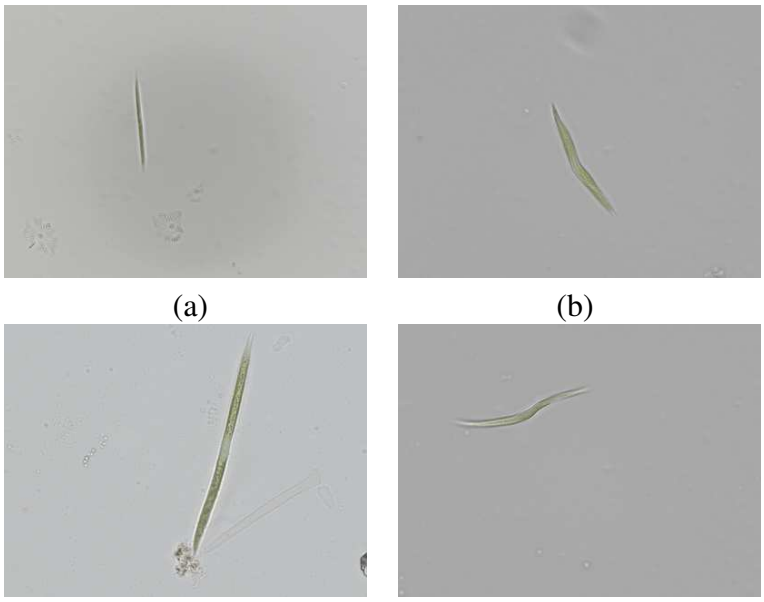
Particularly for green alga images, color is not an appropriate visual property for characterization once various species present intensities similar to green. Furthermore, irregular illumination conditions may result in images with different color patterns. This visual aspect can be observed in the backgrounds of the images in Figure 13, in which shades of grey, light grey and light blue are noted, respectively, in Figures 13(a), 13(b) and 13(c).



(a) (b) (c)

Figure 13 – Color variation in green alga images: different illumination conditions leading to variation in background color between images.

Although no formal definition for texture exists [64], intuitively this visual feature provides information of local intensity patterns, such as smoothness, coarseness and regularity. Popular methods for texture characterization are the co-ocurrence matrix [103] [104] [105], the *Run-Length* [106] [107] and the methods based on fractal dimension [108]. In Figure 12(c), some texture samples are presented for each image region, highlighted by the red, white and blue arrows. Although texture could be applied to identify interesting tissues and components in their organisms, it is not a suitable visual property for accurate description due to the image conditions resulting from the acquisition process. Figure 14 illustrates possible scenarios for algae texture analysis. First, regions associated with alga cells might be blurred, so texture information is not available, as shown in Figure 14(a). Figures 14(b) and 14(c) depict algae from different species that present similar texture patterns. In Figure 14(d), the alga cell presents white areas at the extremities (corners), again suggesting that texture is inappropriate to describe such images.



(a) (b)

(c) (d)

Figure 14 – Textures in green alga images: (a) Blurred alga organisms and compromised texture; (b) and (c) similar texture patterns of two distinct algae genera - *Ankistrodesmus* and *Monoraphidium*; (d) Presence of white areas in the alga cell.

Shape may refer to the silhouette of an object or to spatial information about a target image region which is generally the output of a segmentation process. This visual property is an important feature in human communication and intuitive for the human perception [109]. It is considered the most effective visual property as shape features can reveal a specific object in an image [110]. Figure 12(d) depicts a binary image obtained from a segmentation step, in which the shape silhouette is given by a set of connected white pixels.

Shape has been used for recognizing plants [111], diatoms [112], phytoplankton [113] and other kinds of algae due to its high capability of describing the morphological characteristics of such organisms. The traditional taxonomy of green algae from the Selenastraceae family considers shape as a discriminative feature for distinguishing between algae genus and species. In this family, some species can be differentiated by their shape patterns, while other species share similar shape properties, thus requiring sophisticated extractors to compute more distinctive features. The process of computing features from green alga images is thus an attempt of characterizing the morphological characteristics of these organisms by means of automatic shape-based extractors. Such extractors result from a combination of several shape measures that form a descriptor.

According to *Torres et al.* [114], a descriptor is a pair defined by the feature vector and its associated dissimilarity measure. The dissimilarity measure is usually defined by a distance function that compares two feature vectors. In general, a shape descriptor is a set of statistical measures that attempt to characterize shape similarly to the human intuition. The feature extractor must compute features that are invariant to the following linear geometric transformations:

- Scale: features computed from shapes of different sizes, but from the same category, should be consistent;
- Rotation: the computed features are equivalent for the same shape appearing in different orientations;
- Translation: features computed from a shape appearing in different positions should remain the same;
- Affine transformations: linear mappings from 2D coordinates to another 2D coordinate system that preserve the “straightness” and “parallelism” of lines [115].

It is worth noting that extracting features from alga images, obtaining invariance to scale is not appropriate since alga sizes are relevant for identifying species. Thus, the modeling of green algae descriptors takes into account features that describe numerically size information of alga cells.

Shape descriptors invariant to geometric transformations are consistent with the human perception [110]. A shape might also be robust when dealing with nonlinear deformations, such as:

- occlusion: in which some parts of a shape are hidden by other objects;
- distortion: caused by digitalization errors or inaccurate segmentation of objects and regions;
- noise: noise should not affect the main shape patterns.

Defining appropriate feature extractors and associated dissimilarity measures is highly dependent on the image patterns [99] [7]. Therefore, describing the visual properties of Selenastraceae microalgae requires designing appropriate feature extractors to obtain representative attributes from the segmented images. As aforementioned, the focus is on shape-based features, including some basic geometric measures and other, more sophisticated, approaches that capture complex patterns, as detailed in Chapter 4. The next section presents some dissimilarity measures used to compare images by means of their respective feature vectors.

2.2.4 Dissimilarity measures

The dissimilarity between two data objects (in this case, two images) is a numerical measure that indicates the degree to which they are different [98]. The dissimilarity values are lower for more similar object pairs and higher otherwise. Several such measures have been proposed and popular choices are those based on *metric* functions [116], since they are defined in a metric space. Those measures must satisfy the following axioms for any multidimensional objects $\mathbf{x}$, $\mathbf{y}$ e $\mathbf{z}$ [117]:

- Identity: $d(\mathbf{x},\mathbf{y}) = 0$, if, and only if, $\mathbf{x} = \mathbf{y}$;
- Symmetry: $d(\mathbf{x},\mathbf{y}) = d(\mathbf{y},\mathbf{x})$;
- Positivity: $0 \leq d(\mathbf{x},\mathbf{y}) < \infty$;
- Triangle inequality: $d(\mathbf{x},\mathbf{z}) \leq d(\mathbf{x},\mathbf{y}) + d(\mathbf{y},\mathbf{z})$.

Assuming that $\mathbf{x} = \{x_1,...,x_m\}$ and $\mathbf{y} = \{y_1,...,y_m\}$ are two m-dimensional feature vectors representing two images, the metrics of Minskowski family are given by:

Minkowski distance: also known as the L_∞ norm, it is a general measure that obtains the maximum difference between any attribute of the two objects [98], given by Eq. (2.3):

$$d_M(\mathbf{x},\mathbf{y}) = \left(\sum_{i=1}^{m} |x_i - y_i|^r \right)^{1/r} \tag{2.3}$$

in which $r \in \mathbb{N}$ is a parameter.

Euclidean distance: also known as the L_2 norm, it defines a geometrical spatial placement of all equidistant objects in relation to a reference object, and is given by Eq. (2.4):

$$d_E(\mathbf{x},\mathbf{y}) = \sqrt{\left(\sum_{i=1}^{m} |x_i - y_i|^2\right)} \tag{2.4}$$

City-Block (Manhatan) distance: This is the norm L_1, computed as in Eq. (2.5):

$$d_{CB}(\mathbf{x},\mathbf{y}) = \left(\sum_{i=1}^{m} |x_i - y_i|\right) \tag{2.5}$$

The quadratic distance [118], described by Eq. (2.6), considers the correspondence of an attribute among two objects, as the relationship between attributes of a same object:

$$d_Q(\mathbf{x},\mathbf{y}) = \sqrt{(\mathbf{x}-\mathbf{y})\mathbf{x}^{-1}(\mathbf{x}-\mathbf{y})^T}, \tag{2.6}$$

in which $\mathbf{x} = [x_{ij}]$ is a matrix $m \times m$ and x_{ij} is the similarity coefficient for the dimensions i and j.

A specific case of the Quadratic distance and a generalization of the Euclidean distance is the Mahalanobis distance. This measure is appropriate when the attributes of the objects are somehow correlated, take values on different ranges (different variances) and the data distribution can be approximated by Gaussians [98]. The attribute correlations are described by a covariance matrix obtained prior to computing the distances. The Mahalanobis distance is given by Eq. (2.7):

$$d_\Sigma(\mathbf{x},\mathbf{y}) = (\mathbf{x}-\mathbf{y})\Sigma^{-1}(\mathbf{x}-\mathbf{y})^T \tag{2.7}$$

in which Σ^{-1} is the inverse of the data covariance matrix [64].

The cosine dissimilarity, shown in Eq. (2.8), measures the angle between the directions of both feature vectors:

$$d_{cos}(\mathbf{x},\mathbf{y}) = 1 - \frac{\mathbf{x}\cdot\mathbf{y}}{|\mathbf{x}|\cdot|\mathbf{y}|} \tag{2.8}$$

in which $\mathbf{x}\cdot\mathbf{y} = \sum_{i=1}^{m} x_i y_i$ is the scalar product and $|\mathbf{x}|$ is the cardinality of vector $\mathbf{x}$ ($|\mathbf{x}| = \sqrt{\mathbf{x}\cdot\mathbf{x}}$). Smaller angles indicate higher similarity between the respective data objects.

2.2.5 Image classification

Classification is the process of learning a target function (also known as a classification model) that maps each instance from a data set to one of a predefined set of categories [99]. Classification models can be devised to distinguish between images of different classes or to

predict the class label of unknown images. Classifiers employ a systematic methodology to build classification models from an input image set, consisting of a learning algorithm that identifies a model that best fits the relationships between image attributes and the class labels of the input data [98].

The learning algorithm should build models with a good generalization capability by predicting accurately the class labels of images. A general methodology for building a classifier requires a training set, which consists of instances with known class labels. The training set is the main reference for the learning algorithm to build a classification model. After, an instance set with unknown class labels, also known as the test set, is taken as input to the classification model to evaluate its performance.

There are several classification techniques that rely on distinct mathematical formulations or methodologies for training and testing. The choice of classifier is a difficult task since it depends on the problem characteristics. It is a common practice when addressing a classification problem to perform some experimentation with several classifiers, considering also the data attributes and class label information [119]. Support vector machines [120] [121] and artificial neural networks [122] [123] [124] [28] have been widely applied in pattern recognition tasks on biological data.

Artificial neural networks (ANN) are a family of mathematical models inspired by biological neural networks which are defined by many artificial neurons correlated together in accordance with an explicit network architecture. In a supervised learning approach, the training step of an ANN classifier consists of processing the training instances one at a time and learning by comparing the predicted classes with their real labels. If an instance has been mistakenly predicted the network parameters are immediately updated by means of a feedback into the network. The training step of ANNs employs some form of gradient descent scheme, since many iterations can be performed until the network obtains the set of parameters that best fits the training set instances and their class labels.

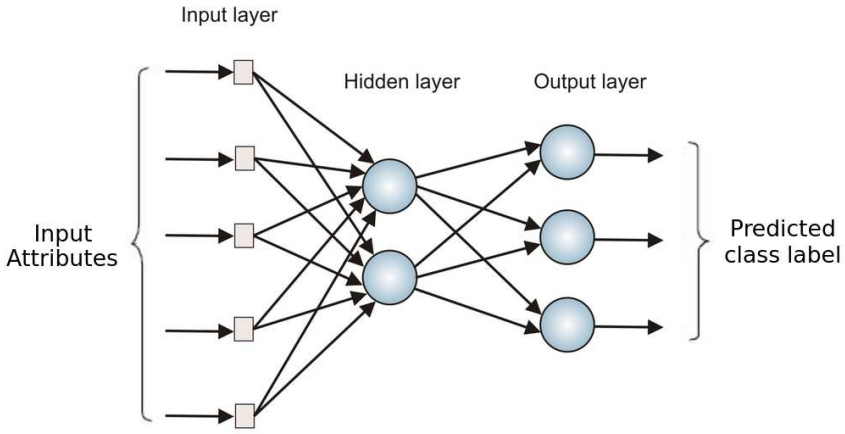


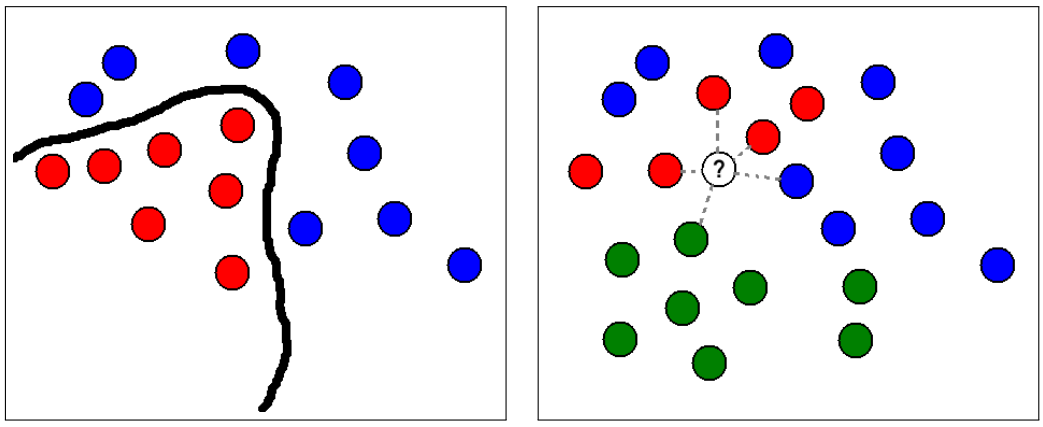
Figure 15 – Artificial Neural Network based on Multilayer Perceptron.

A popular ANN architecture is the Multilayer Perceptron (MLP) which represents the ar-

tificial neurons as nodes and their connection by weighted links. The input nodes form an input layer and are represented by the attributes of the input instances, while the output nodes constitute the output layer used to represent the model prediction. The network may contain several intermediate layers (also called hidden layers) in which the nodes in one layer are connected only to the nodes in the next layer. Figure 15 depicts the architecture of a MLP by showing its constituting layers and node organization.

The project of an ANN-based classifier requires defining the number of hidden layers, as well as the number of its constituting nodes. Furthermore, some settings should also be defined according to the data characteristics concerning the Backpropagation algorithm that trains the ANN, such as the choice of a differentiable activation function (sigmoid or hyperbolic tangent) and the learning rate to update network weights. Additional information on ANNs can be found elsewhere [125] [126] [127].

Support Vector Machine (SVM) is a classification technique successfully employed in many practical applications, as it can handle data of varied patterns or presenting high dimensionality [98]. SVM represents a training set as a high-dimensional space in which instances are represented as points. Considering a binary classification problem, it attempts to find a decision boundary based on a maximal margin hyperplane that separates linearly or nonlinearly the training set into two classes. The training step attempts to compute such hyperplane by maximizing its distance to the nearest data point on each side. Non-linear classification problems require determining a decision boundary using the "kernel trick", which implicitly maps the training instances to a high-dimensional space. Figure 16(a) illustrates a nonlinear decision boundary determined after training an SVM classifier, which discriminates the two categories of instances.



(a) (b)

Figure 16 – Illustration of learning models: (a) nonlinear decision boundary obtained from a Support Vector Machine classifier; (b) K-Nearest neighbors ($K = 5$), in which the non-labeled instance '?' will assigned the red label.

One of the simplest classification technique is K-Nearest Neighbors. It assigns to a test instance the most frequent class label from its K most similar training instances, as given by an appropriate dissimilarity measure applied to their feature vectors. The main challenge when

adopting a classifier based on the nearest neighbors principle is to define the value of K, since it depends on the number of class labels and on class balance. Figure 16(b) shows the label assignment process after computing the distances for the unknown instance '?' and verifying that the three most similar samples belong to the red class.

Decision trees are a widely applied classification technique and useful in problems in which the feature space can be partitioned by considering satisfying conditions over the features [128] [129]. Tree leaves represent class labels, internal nodes are attribute test conditions to separate instances with different characteristics and branches represent conjunctions of features that lead to the class labels. The training step consists of building the tree by determining a key attribute which partitions the attributes set at each node (root and internal nodes). A test instance is classified starting from the root node, applying the test condition to the instance and following the appropriate branch based on the outcome. This decision leads to another internal node, where the next test condition is applied, or to a leaf node. Then, the class label associated with the leaf node is assigned to the test instance. Figure 17 illustrates a decision tree for classifying generic species of an alga family, in which the leaf nodes are associated with the species.

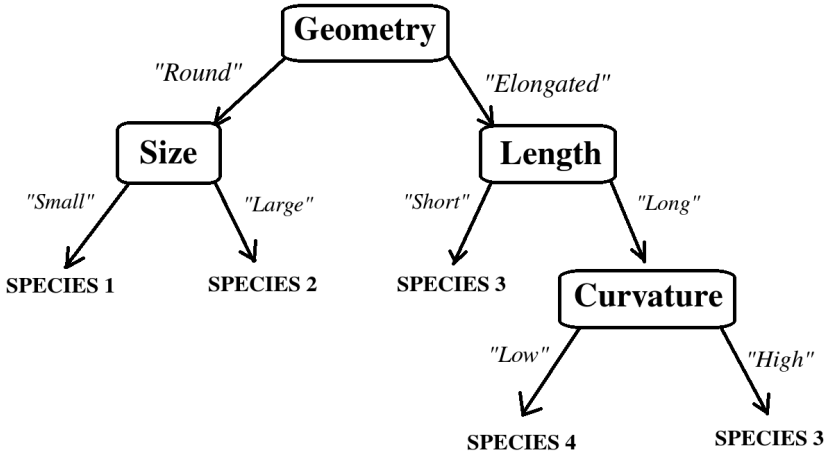


Figure 17 – Decision tree for classifying species of a generic alga family.

Decision tree classifiers can be differentiated according to the strategy adopted to build the tree and to model the nodes [130] [131]. The ID3 algorithm is a popular approach for building decision trees [132] and has been successfully applied to study public health [133], medicine [134] and genetics [135]. ID3 is an iterative algorithm that analyzes every unused attribute from the attributes set of the training data and calculates its entropy (or other information gain measure), and picks the attribute with the lowest entropy value (or highest information gain). The training data is split by the selected attribute to produce subsets. This process continues to recurse on each subset, considering only attributes not yet considered.

Evaluating classifier performance

The correctness of classification models can be evaluated by counting the number of test instances that are correctly and incorrectly predicted by the model. Consider a binary classification problem (two class labels) that produces discrete output values. Each instance is thus assigned to one class of positive (P) or negative (N) class labels. Given a classifier and an instance, there are four possible outcomes. A positive instance classified as positive is counted as a true positive (TP). A positive instance predicted as negative by the classifier is counted as a false negative (FN). A negative instance correctly classified as negative is counted as a true negative (TN), and if an instance is classified as positive, but its real class is negative, a false positive (FP) is counted. These counts generate a 2×2 confusion matrix [136], which is used to derive some convenient metrics to summarize quantitatively and qualitatively the model's performance by means of single values [98].

From the classifier confusion matrix, the true positive rate TPR of a classifier, also called hit rate or recall, is estimated as:

$$TPR = \frac{TP}{TP+FN}, \tag{2.9}$$

and is defined as the rate of positive examples predicted correctly. The false positive rate (FPR), popularly known as the "false alarm" rate, is estimated as:

$$FPR = \frac{FP}{TN+FN}, \tag{2.10}$$

and refers to the rate of negative examples predicted as positive. Accuracy (Acc) measures the correct prediction of positive and negative instances in the test set:

$$Acc = \frac{TP+TN}{TP+FP+FN+TN}. \tag{2.11}$$

However, accuracy is not a suitable measure on imbalanced datasets. In this case, precision and recall are more appropriate for a fair evaluation. Precision (PRC) measures the correct predictions of positive instances from all the positive predictions:

$$PRC = \frac{TP}{TP+FP}. \tag{2.12}$$

Precision and recall can be combined into another metric, known as F_1-Measure:

$$F_1 = \frac{2 \times TP}{2 \times TP+FP+FN}. \tag{2.13}$$

High F_1 scores express high values for both precision and recall.

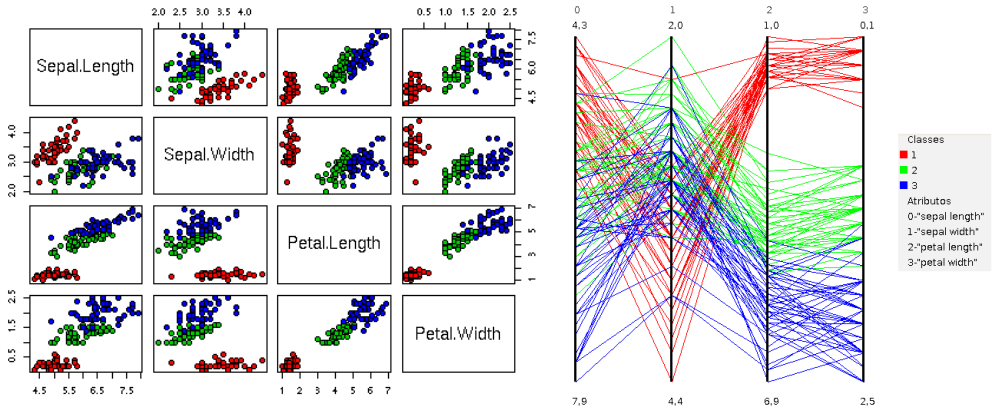
Classifiers should ideally attain the highest classification performance according to

some of the described metrics. For that purpose, selecting the classification and learning models, defining representative training sets and adjusting classifier parameters is a complex task that is highly dependent on data patterns [35] [34]. Visualizations allow a direct interaction with the user and provide immediate feedback, as well as user steering, which is difficult to achieve in traditional data analysis.

2.2.6 Multidimensional visualization

Information visualization can contribute to data analysis by generating interactive visual representations that demand understanding and extracting information from relevant patterns in datasets [137] [138]. The use of visualizations is encouraged in data analysis since humans are trained to perceive interesting patterns in graphical representations. Thus, visual representations of data can significantly contribute to successful knowledge extraction from data, particularly in exploratory tasks [139].

Visualization techniques can be categorized as attribute-based or observation based [140]. Techniques that rely on attributes generate visualizations in which the data atributes (dimensions) are explicitly shown. Examples of traditional attribute-based visualizations are Scatterplots and Scatterplot Matrices [141] and Parallel Coordinates [142]. Figure 18 shows visualizations of the *Iris* dataset [1], which describe 150 data instances (observations) by means of 4 attributes. Figure 18(a) illustrates a Scatterplot Matrix that shows scatterplot views of all attribute pairs arranged in a matrix structure. Figure 18(b) shows a Parallel Coordinates view of the same data, in which attributes are mapped to parallel vertical axes and each observation is represented as a polyline.



(a) (b)

Figure 18 – Visualization of data attributes in data set *Iris* [1]: (a) Scatterplot Matrix; (b) Parallel Coordinates.

Although these classical visualizations are useful to explore low dimensional data sets, users begin to face interpretation difficulties as the number of data dimensions increases [4]. Observation-based visualization techniques, also known as point-placement visualizations, create graphical layouts by placing visual elements (circles, points, or spheres) representing individual data instances in a visualization space (a one, two or three dimensional space) [143]. The

visual output is a point cloud in which the relative positions of points reflect some relationship amongst the corresponding data instances in the original space, such as pairwise dissimilarities or neighborhoods. The data attributes are not explicitly represented. In handling large data sets, many points need to be simultaneously represented in a limited visual space and cluttering may impair interpretation, e.g., the identification of groups of correlated instances [144].
Similarity trees are an alternative point placement strategy that can reduce the clutter and provides an intuitive approach to represent similarity relations between data instances as a hierarchy, e.g., for an image dataset they can provide a graphical representation of the similarities between the image feature vectors. Visually, a similarity tree shows points placed in branches and sub-branches, where points representing more similar data instances are placed in the same or in closer branches, and points representing more dissimilar instances are placed in branches farther apart. As such, similarities are organized as a hierarchy that facilitates interpretation, as compared with multidimensional projections.

A particular case of similarity trees are the phylogenetic trees. According to *Leliaert et al.* [145], phylogeny can be defined as the study of the evolution process of groups of species and their relationships. Phylogenetic trees present a topology that emphasizes the degrees of relationship between members, in which the leaf nodes represent the species, while the internal nodes express the hypothetical ancestors. Branch length expresses the evolutionary distance between each member in a group [146]. In data visualization, the use of phylogenetic trees allows to preserve the neighborhood relationships defined in the original space and enables to identify the density of groups of similar points in a data set by analyzing the tree branches [2].

A popular algorithm for building phylogenetic trees is the *Neighbor Joining* (NJ) [147]. The NJ algorithm builds a phylogenetic tree by replacing the concept of ancestry by a virtual object node with combined dissimilarity. The tree is created based on the property that similar nodes (instances that share common properties) are assigned to the same branch, in a bottom-up strategy. Therefore, a node is assigned as an ancestor of another node when they share similar content. The reduced tree depth yields rational use of screen space, and a useful hierarchical interpretation that reveals both local and global similarity relationships.

The Neighbor Joining heuristic builds a unrooted tree from a distance matrix computed from the data. The rationale is to search for the closest pair of instances, which are connected by an internal node into a binary subtree. Starting from a star-like tree, described as n leaves connected to a single central node, the NJ algorithm can be described by the following steps:

1. Choose the pair of instances that provides the smallest branch length sum S_{ij}, as shown in Eq. (2.14):

$$S_{ij} = \frac{1}{2(n-2)} \sum_{k \neq i,j}^{n} (D_{ik} + D_{jk}) + \frac{1}{2} D_{ij} + \frac{1}{n-2} \sum_{(k,l \neq i,j) \wedge (k<l)}^{n} D_{kl} \tag{2.14}$$

in which D_{ij} is the respective value in the distance matrix, k represents the instances i and

j, and n refers to the number of data instances.

2. Create a virtual node X and add it to the tree, having i and j as children and connected to their common ancestor.

3. Compute the length of the new branches according to the *Fitch-Margoliash* method, given by Eq. (2.15):

$$L_{iX} = \frac{D_{ij} + D_{iz} - D_{jz}}{2} \qquad L_{jX} = \frac{D_{ij} + D_{jz} - D_{iz}}{2} \tag{2.15}$$

in which z represents the tree instances, except for instances i and j. Distances D_{iz} and D_{jz} are the respective distances from instances i and j to all the remaining nodes, given by Eq. (2.16):

$$D_{iz} = \frac{\sum_{(k \neq j)}^{n} D_{ik}}{n-2} \qquad D_{jz} = \frac{\sum_{(k \neq i)}^{n} D_{jk}}{n-2} \tag{2.16}$$

4. Update the distance matrix to embed the new virtual node X representing the pair of neighbor instances i and j and compute the distance from X to the remaining instances, as in Eq. (2.17):

$$D_{X,k} = \frac{D_{ik} + D_{jk}}{2} \tag{2.17}$$

in which $k \leq n$, except for i and j.

5. Repeat these steps until there are only two nodes remaining in the distance matrix.

In an NJ-tree visualization, n instances are graphically represented as n leaf nodes and $n-2$ internal nodes. Given as input a matrix of pairwise distances for the dataset instances, the NJ algorithm is executed and a technique is applied to draw a visual representation of the tree with its nodes and branches. Figure 19 presents an NJ-tree visualization of the Corel500 dataset [2]. The colors map images from the same category (class), i.e., groups of images with similar characteristics. A visual analysis shows that the visualization can satisfactorily group the categories of images. The NJ-tree visualization preserves the data similarity relations due to the embedded hierarchy built using their pairwise distances.

Cuadros et al. [39] studied the applicability of NJ-tree visualizations on multidimensional data from textual documents. The tree-based visualization was devised in a way that leaf nodes represent the documents and the edge lengths express the similarity among documents. Experiments validated the NJ-based visualization, which created a hierarchical structure preserving the neighborhood relations. Furthermore, a visual analysis of the NJ-visualization showed that similar documents were placed in the same tree branches.

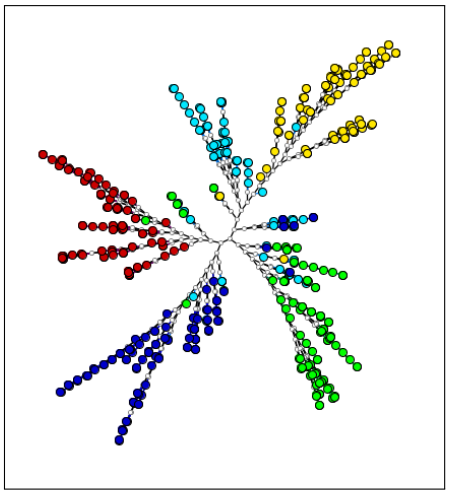


Figure 19 – NJ-tree point placement visualization of the Corel500 data set [2].

NJ visualizations make a better usage of the visual space than multidimensional projections, because the tree-based representations can be expanded to reduce significantly or to remove node overlapping. Another advantage is their ability of preserving the distance relations in the original space both in a global aspect, when wide neighborhoods are considered, or locally, when considering close neighborhoods.

The computational complexity of the NJ algorithm is $O(n^3)$, because a $n \times n$ distance matrix is analyzed to select and combine the nearest neighbor nodes, and the process that searches the neighbors is performed $n-2$ times at each iteration. In addition, the NJ algorithm has an undesired drawback of generating a large number of virtual nodes. Handling several virtual nodes is computationally costly and leads to limitations when data instances are shown simultaneously, influencing the analysis of global data relations [148]. Researchers introduced modified formulations of the original NJ formulation aiming at a better computational performance or at improving the quality of tree layout, such as the *Fast Neighbor Joining* (FNJ) [149], the *Rapid Neighbor Joining* (RaNJ) [150] and the Promoting Neighbor Joining (PNJ) [2].

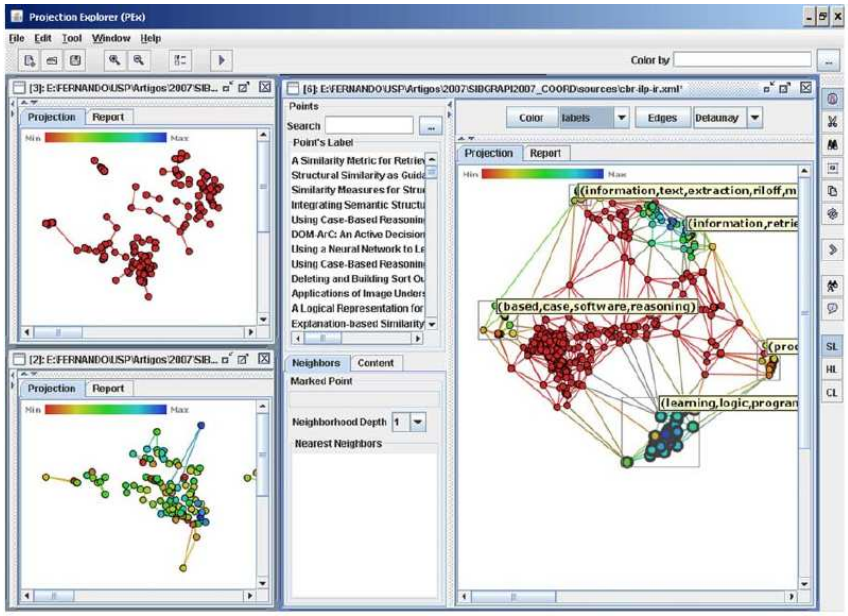
2.2.7 Interactive visual data mining

Visual data mining has shown great potential in exploratory analysis tasks over large high-dimensional datasets [151]. The underlying rationale is to employ visualizations in traditional data mining tasks in which the similarity relations between data instances and dimension correlations are less well known [152]. This process may lead to the visual discovery of relevant patterns in data or provide some guidance for the application of other data mining and analytics techniques. Generally, visual data mining allows users to gain deeper understanding of the relevant data patterns and relationships.

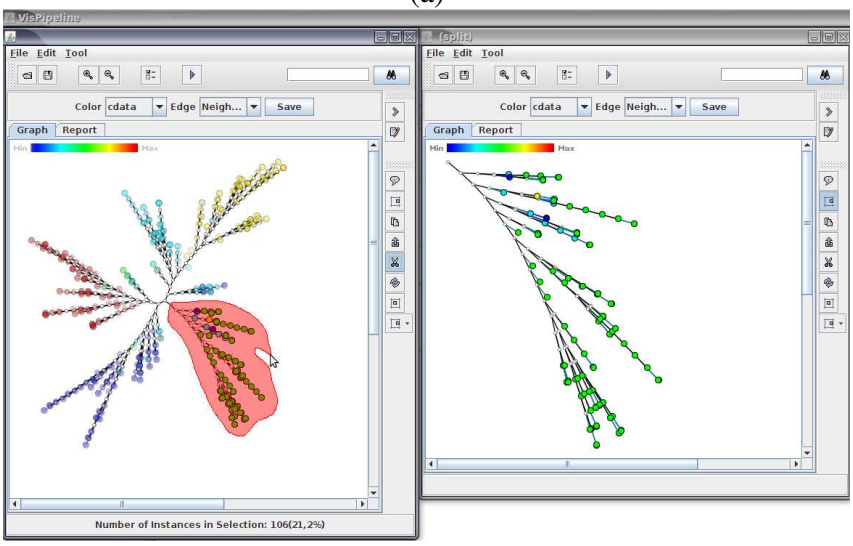
Visual data mining processes can benefit from user participation during data analyzis, exploration and mining tasks [153]. Several authors argue in favor of including end users into data mining processes, because it aggregates flexibility, creativity and knowledge allied to the

powerful processing of current computers [140] [154] [155] [156].

Interaction tools are required to enable users to handle the visualizations and modify the visual mappings according to the data exploration goals. Figure 20 shows two examples of user interaction techniques. Figure 20(a) illustrates the Coordinated Multiple Views [3], in which multiple data views are shown and interaction actions executed over one of the views are reflected in the others. Thus, the user interacts with multiple coupled data representations of the same or related datasets. Figure 20(b) depicts the filtering technique that consists of selecting a subset of points in the visualization, e.g., for further detailed analysis. In this case, a tree branch was manually selected and presented in another visualization [4].



(a)



(b)

Figure 20 – Interactive tools: (a) Coordinated Multiple Views [3]; (b) Interactive filtering of samples using the selection tool in *VisPipeline* [4].

Visual exploration of large data sets has been used as a complementary technique to

data mining to obtain additional information about the data. *Keim* [35] characterizes the visual exploration as a robust process capable of handling diverse kinds of data, including those which are heterogeneous or noisy. Moreover, this process does not require advanced knowledge about mining and learning algorithms embedded in the application. Thus, visual exploration can make data analysis tasks more effective in scenarios where mining algorithms are likely to fail.

According to *Keim et al.* [37], visual exploration is a three-fold process: overview first, zoom and filter, and then details-on-demand. In the *overview first*, users employ visualization techniques to get an overview of the data to identify interesting patterns, focusing on one or multiple patterns. Then, in the *zoom* and *filtering* step, the user needs to drill-down and access details by visualizing subsets, but maintaining the overview visualization. Finally, *details-on-demand* refers to searching for details in the visualizations until they obtain sufficient information. It is important to note that visualizations and interaction tools are employed in all three steps but also bridging the gaps between the steps.

A previously studied task in visual data mining is visual image classification. Specifically, *Paiva et al.* [2] introduced a visual classification methodology for image collections of different natures. Visual classification is a user-assisted process that employs exploratory visualization to support classification tasks in scenarios where the relevant relations between data objects are assumed to be unknown. This process relies on visualizations capable of presenting users with a meaningful graphical metaphor capable of revealing relevant data patterns. Inserting the user into the classification loop requires some interactive tools that allow introducing user feedback based on previous knowledge or preliminary analysis. Thus, they are immediately exposed to the classification results, including false positives, false negatives, mismatches and outliers, and also to input information into the process. Experiments to validate the visual classification process were performed on real-world image sets, which results indicated that the classifications accuracy improved as iterations succeed.

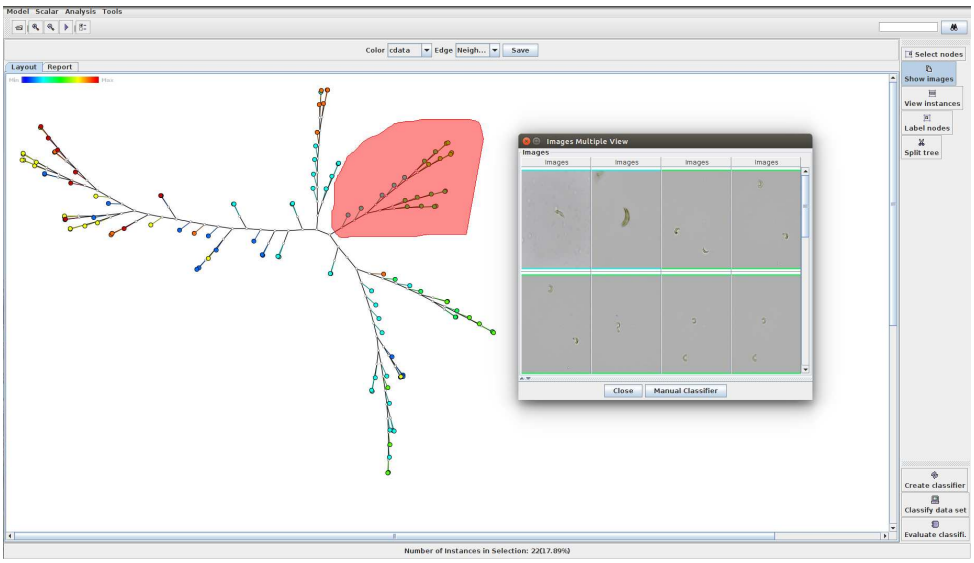


Figure 21 – Layout of the Visual Classification System [2].

Furthermore, *Paiva et al.* developed the Visual Classification System (VCS), a framework that implements multidimensional visualizations (projections and phylogenetic trees) and classification techniques (KNN and SVM), as the required resources to enable user-assisted data exploration and classification. VCS incorporates interactive techniques to support users adjusting classifier settings and interpreting visualization and classification results. One of its functionalities is a tool for labeling data instances using visualizations, which allows users to select representative instances for training sets. Another tool is the Class Matching, that supports user analysis of classification results and guide them through taking better decisions that improve the classification process.

In this work, VCS has been modified to support biologists in their algae classification tasks, as described in Chapter 5. Several functionalities have been incorporated into VCS to facilitate its usage by the biologists, who provided feedback for the development of the functionalities in the current prototype.

2.3 Related work

This section presents a literature survey on classification of alga species using image processing and pattern recognition techniques. As we are not aware of previous studies concerning the automatic classification of Selenastraceae, some methodologies targeted at other alga species are described. Although the intensity patterns and image conditions of Selenastraceae images are different than those seen in other alga images, existing solutions can certainly influence the definition of specific approaches. It is worth to emphasize, whenever applicable, important aspects to this work, such as the image segmentation strategy, the approach for algae characterization, the classifier modeling and the classification accuracies achieved.

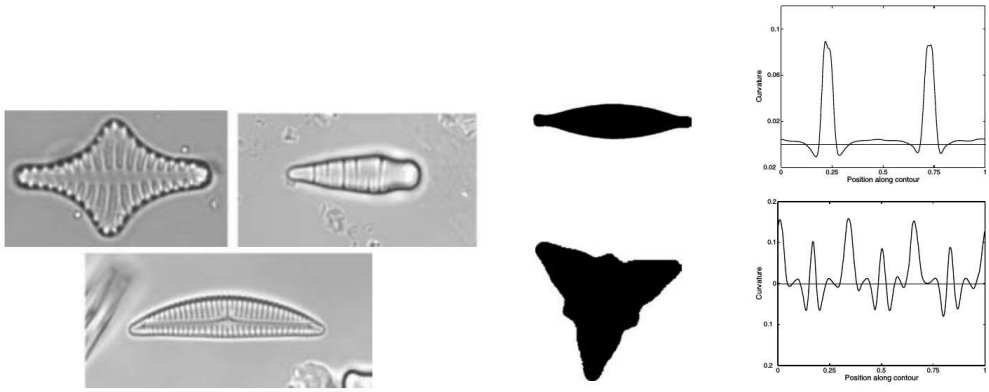
The classification methods for biological and alga images are categorized based whether they are automatic or adopt a visual exploration process.

2.3.1 Classification of alga species in digital images

Automatic methodologies for algae classification ideally only require users to indicate the input image before outputting a label associated with a species. In most cases, a single system with a Graphical User Interface (GUI) has been developed implementing the image processing, feature extraction and classification methods, meant to be used in multiple knowledge domains.

Jalba et al. [5] proposed a method for automated identification of diatoms using shape-based features extracted with multi-scale mathematical morphology techniques, as illustrated in Figure 22. The diatom shape is represented in a curvature scale space, which consists of applying an adaptive smoothing to the shape contour and computing curvature information at each iteration. After transforming the data in the curvature scale space to the scale space, a feature selection step based on the mean-shift algorithm removes noisy and redundant information that

may arise in the scale space. The final diatom descriptor is invariant under translation, rotation and scale. Experiments were performed in two large sets using a C4.5 decision tree [157] and the K-nearest neighbors classifier. Two combinations of feature vectors based on curvature measures were considered. Authors reported that the best set of features achieved 84% identification accuracy and the automated methodology generated similar results when compared with the same process conducted by human experts.

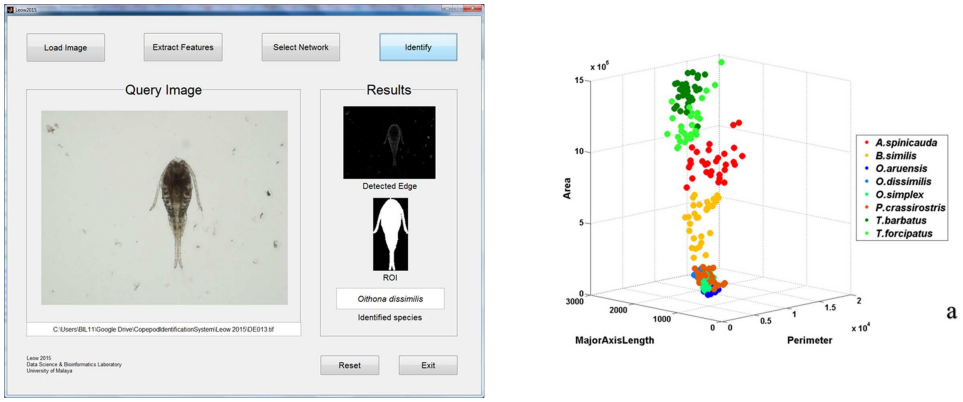


(a) (b)

Figure 22 – Jalba's method [5] for identifying diatoms: (a) Samples of diatoms cells; (b) binary images of diatoms and their respective curvature plots.

Leow et al. [6] developed an automated technique to extract morphological characteristics and to identify species of copepods' from digital microscope images. Images based on dorsal images of copepods are segmented via binary thresholding after applying a median filtering to the original image. Shape-based geometric features, such as area, major axis, perimeter and convex area are computed over a region of interest obtained from the segmented image. An ANN with two hidden layers and conjugate gradient backpropagation was used to classify the extracted features into eight classes (species). A total of 240 sample images were used in the experiments, with 60% of the dataset used for network training and the remaining 40% used as a test set. The overall classification accuracy obtained by the method is 93.13%. Figure 23(a) depicts the system's GUI for automated identification of copepods, in which a user can handle images and perform operations such as segmentation and feature extraction, and classify samples using the ANN-based classifier. Figure 23(b) presents a graphical comparison using three out of the eight features used to describe the shapes of copepods' species.

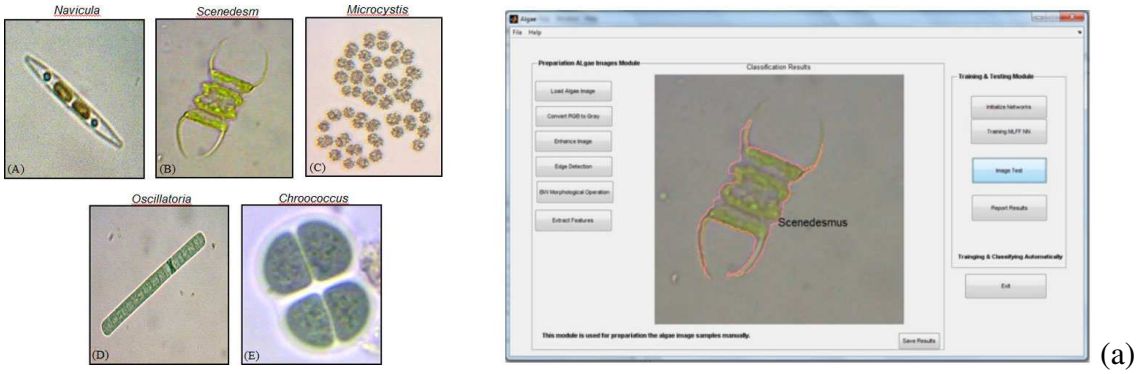
Mosleh et al. [7] developed an automated freshwater algae detection system to detecting, recognizing, and identifying alga genera from the divisions Bacillariophyta, Chlorophyta and Cyanobacteria. First, image preprocessing was applied to improve contrast and remove noise from the original microscope images. The obtained images are then segmented using the Canny edge detector algorithm, resulting in a binary image with the algae and its boundaries. Shape features, such as area, perimeter, minor and major axes are then computed. Moreover, Fourier analysis with principal component analysis (PCA) was applied to the enhanced image



(a) (b)

Figure 23 – Leow's method [6] for identifying copepods species: (a) a Graphical User Interface (GUI) of the system; (b) Scatterplot of a subset of the basic geometric shape features.

to obtain texture information. An ANN is used for classification, with a feed-forward multilayer perceptron network with backpropagation trained on 50 sample images. Comparing the manual classification results with the automatic approach, 93 images of selected freshwater algae genera were correctly identified from a total of 100 tested images. The training step takes around 5 minutes and the time required to classify each test image varied from 1 to 1.5 minutes. Figure 24(a) illustrates some alga samples for the identification process and Figure 24(b) shows the prototype interface implementing the image processing, feature extraction and classification techniques.



(a)

(b)

Figure 24 – Mosleh's method [7] for identifying algae genera from the divisions Bacillariophyta, Chlorophyta and Cyanobacteria: (a) Algae from the Bacillariophyta, Chlorophyta and Cyanobacteria divisons; (b) the GUI for automated identification.

Coltelli et al. [8] proposed an automatic methodology for real time identification and enumeration of microalgae using image processing and pattern recognition techniques. The methodology employs a mixture of multivariate Gaussians computed from color samples of target algae to detect and recognize in-focus and out-focus microalga cells that can appear in a microscope slide. The segmentation approach is contour-based and detects boundaries of microalgae and unwanted objects (bacteria, particles and dead cells), which are disregarded during the recognition process. Some morphological and densitometric features based on shape are extracted,

such as the center of gravity coordinates, area, Feret diameters, extinction, among others, totaling 94 numeric features. For the sake of performance, only a subset of selected features are input to a Self-Organizing Map (SOM) [158] for grouping similar alga species. The training set consists of 16,161 samples represented as feature vectors and SOM organizes them into homogeneous groups. Other 53,869 images were used as test set and the clustering approach attained 98.6% of accuracy, which is higher than the values achieved by previous identification and classification methods, and very close to those obtained by a phycology expert. Figure 25 describes some steps of Coltelli's methodology: the process of computing shape features and the SOM employed for classification.

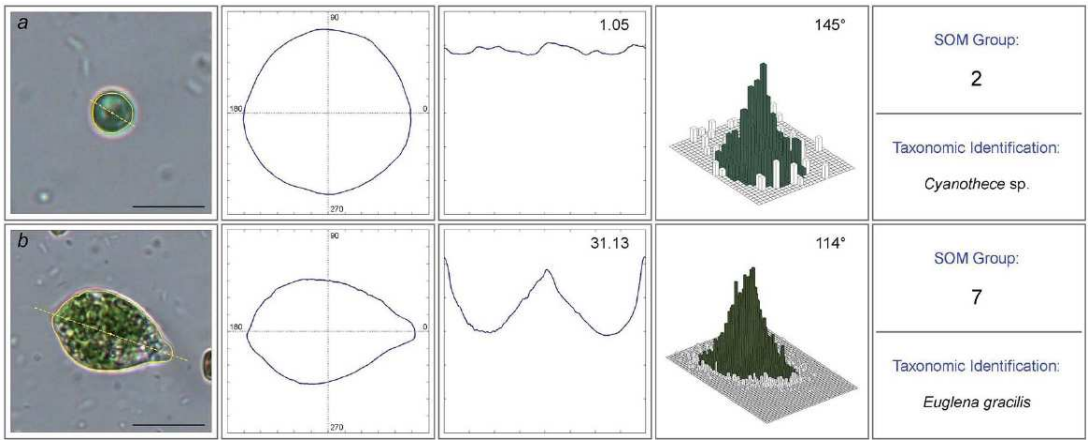
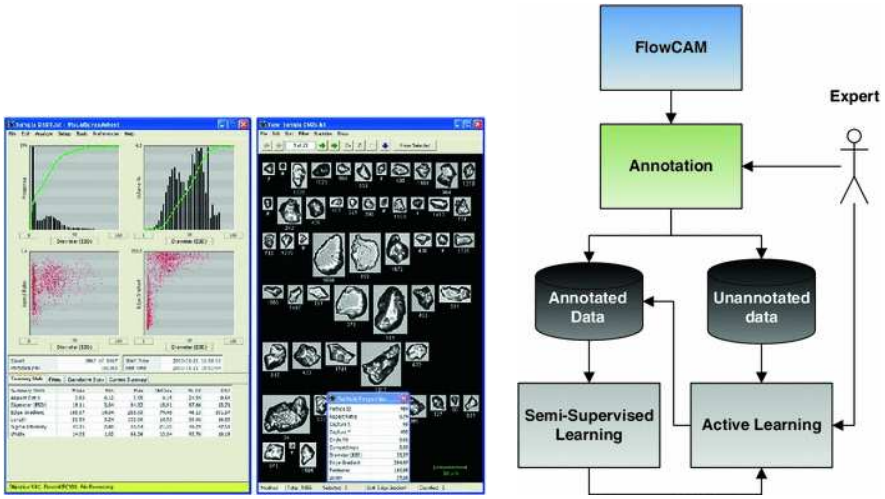


Figure 25 – Coltelli's method [8] for real-time identification and enumeration of microalgae in images.

Drews Jr. [9] proposed a method to classify microalga species based on an active learning strategy [159] and a semi-supervised approach. First, the microscope images are segmented by means of guided segmentation using the software FlowCAM, which is also employed for extracting basic shape geometric features, as shown in Figure 26(a). Drew's classification methodology is naturally iterative, consisting of the two major steps depicted in Figure 26(b): a process for labeling images using active learning with user participation and the algae identification using a classification model trained from these labeled images. Starting from an initial clustering, the first step presents some images for user labeling, as selected by the active learning algorithm, which analyzes the non-labeled images searching for those which will add the higher information gain to the classification model. The user informs their labels and the current training set is updated with the labeled images. The Expectation-Maximization [160] algorithm is executed on the updated training set and the clusters obtained define the classification model. A classification is performed on a test set and the user checks the classification results. If they are not satisfactory, new images are presented to the user for labeling and added to the previous training set. A new classification model is then obtained with EM and the classification process is performed on the test set. Experiments were performed to validate the proposed approach, which achieved near 92% accuracy, while accuracy results obtained by humans remained between 67% and 83%.



(a) (b)

Figure 26 – Drews's methodology [9] for microalgae classification: (a) FlowCAM; (b) Flowchart of the proposed semi-supervised classification approach based on active learning.

2.3.2 Visual exploration of biological data

Visual exploration processes have been adopted in several applications that employ mining tasks or apply data analysis to infer meaningful knowledge [161] [162] [139]. The major motivation for employing a visual exploration process in the classification of the green alga species is the possibility of the biologist contributing to the task. Being a difficult task, the taxonomical identification of Selenastraceae, with a computational user-assisted classification methodology appears as an interesting alternative. It can support the biologists to iteratively refine a classification and improve its effectiveness. Such a strategy has been employed in similar scenarios, as discussed next.

Fernstad et al. [10] designed the MicrobiVis environment to support users in the visual exploration and interactive analysis of high-dimensional data relative to microbiomic population. The proposed system features tools such as dimensionality reduction, exploratory data analysis and standard quality metrics for the microbiology domain. This software employs Scatterplots and Parallel Coordinates to convey the patterns in microbiological data, as well as phylogenetic trees to represent the hierarchical relations among taxon units. MicrobiVis includes a ranking algorithm that allows the biologist to identify anomalies and interesting structures. Such system was validated considering the positive feedback provided by biologists concerning its usability and potential for analysis on DNA-based data on microbes. Figure 27 presents a panel showing a filtering applied to a Parallel Coordinates view and its linking to other visualizations, such as another Parallel Coordinates view and a phylogenetic tree.

Hasenauer et al. [11] proposed a visual analytics approach to support the analysis of models of heterogeneous cell populations, such as cancer and stem cells. In their work, cells are described by biological attributes and other predetermined parameters. The proposed method combines Parallel Coordinates plots, used for evaluating the high dimensional dependencies,

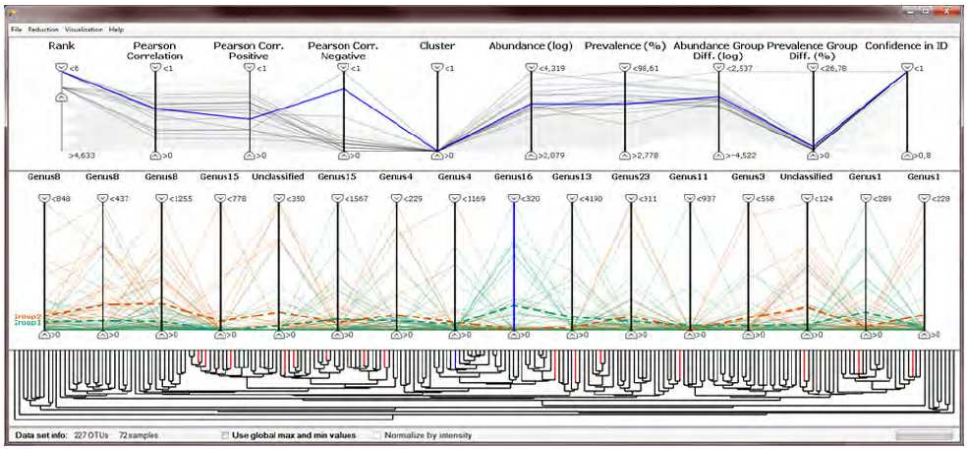


Figure 27 – MicrobiVis [10]: a panel showing a filtering operation in a Parallel Coordinates and its linking to other visualizations.

and Support Vector Machines, for the quantification of bio-markers quality. Parallel Coordinates allow users to select the most relevant bio-markers by interpreting the visualization, which also reveals interesting patterns between the dimensions. A nonlinear SVM is employed for a quantitative assessment of the model's representativeness considering the selected candidate bio-markers. Experiments were conducted aiming to study the influence of model's parameters regarding the heterogeneity of the cells population. Their studies suggested that the employment of Parallel Coordinates allows user to select of potential parameters by interpreting patterns of cell populations in the visualization. Figure 28 presents a workflow of the Hasenauers' methodology and the integration of each of the modules: data processing, modeling and parameter estimations, and the analysis of cell population models, which contains the modules for the SVM classifier (and regression) and the Parallel Coordinates.

Francisco et al. [12] introduced a software called PHYLOViZ, which allows the integrated analysis of sequence-based typing methods, including data generated from whole genome sequences, and associated epidemiological data. Figure 29 shows the software layout, which includes an expansion module based on a Minimum Spanning Tree for visualizing possible evolutionary relationships between isolated data instances. The visualization of typing data profiles and isolate data allows users to query and filter the data using regular expressions or simply pointing and clicking. A usage case on a data set from *Streptococcus pneumoniae* has shown some global and local visualizations based on the MST, in which sequencing types and penicillin resistance information were used as isolates. In the proposed visual exploration process, users can instantly display how the penicillin resistance varies throughout the instances. The results can be displayed as an annotated graph overlaying the query results of any other epidemiological data available.

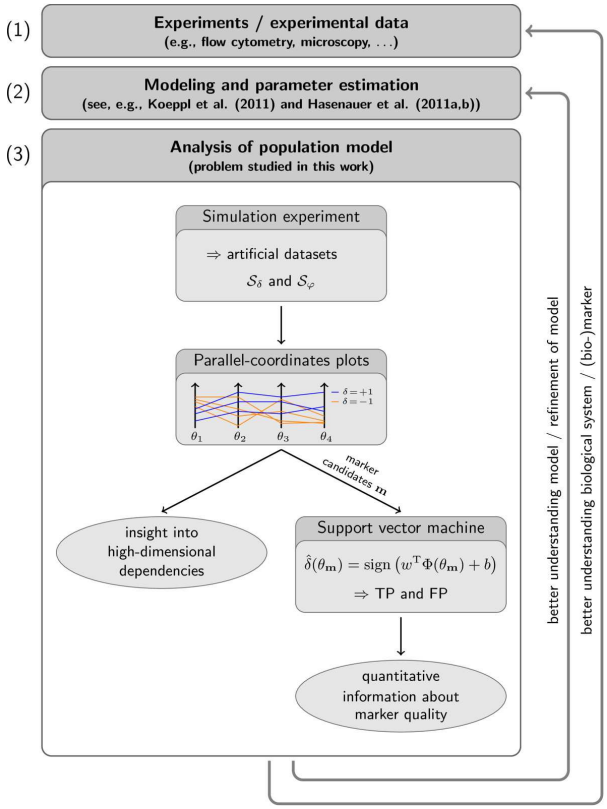


Figure 28 – Flowchart of Hasenauer's visual analytics approach [11].

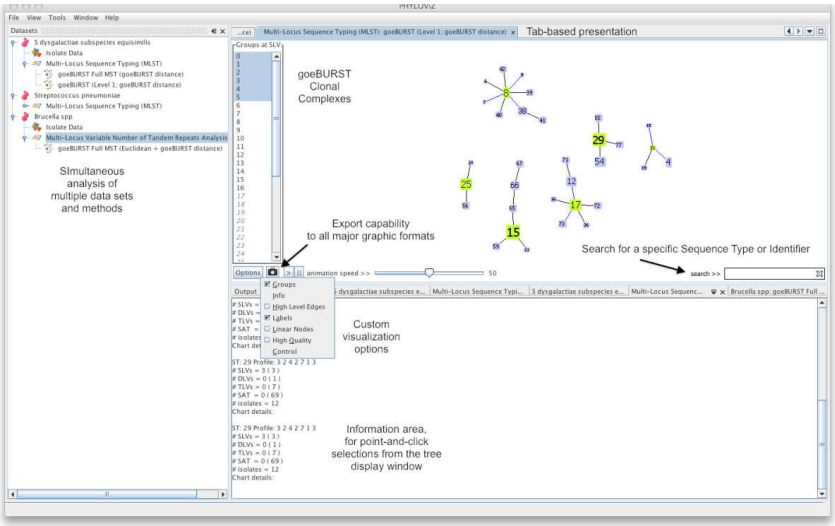


Figure 29 – The layout of PHYLOViZ [12]: several panels depict different functionalities to support users to elaborate queries and interpret results in visualizations based on minimum spanning trees.

2.4 Final considerations

This chapter presented a discussion on the limitations of current practices for the taxonomical identification of green microalga species faced by biologists. Moreover, the microalga species of family Selenastraceae were described, as well as their main morphological properties used to characterize the species. This information is essential to design the image preprocessing

and segmentation, pattern recognition and visual analysis techniques that compose the proposed methodology for the taxonomical classification of Selenastraceae algae.

The fundamentals and concepts on image processing, feature extraction, automatic classification and multidimensional visualization have been described using as reference related work that applies computational strategies to analyze biological data or to classify other alga species. Literature shows that most solutions are limited to particular species. So, this research is aimed at filling this gap regarding a particular type of microalgae, i.e., to classify digital images of the Selenastraceae microalgae at the species level, using the referred techniques.

The next chapter describes the proposed segmentation methodologies for obtaining representative alga shapes from microscope images.

CHAPTER

3

SEGMENTATION OF GREEN MICROALGA IMAGES

Image segmentation is a difficult computational task which consists of subdividing the image domain in its constituting regions according to a criterion based on similarity or boundaries. The quality of the segmentation directly affects further processing steps, such as the extraction of representative features. Defining an appropriate segmentation technique is highly dependent on the application and requires a solid knowledge on specific image properties such as brightness, noise, texture and contrast.

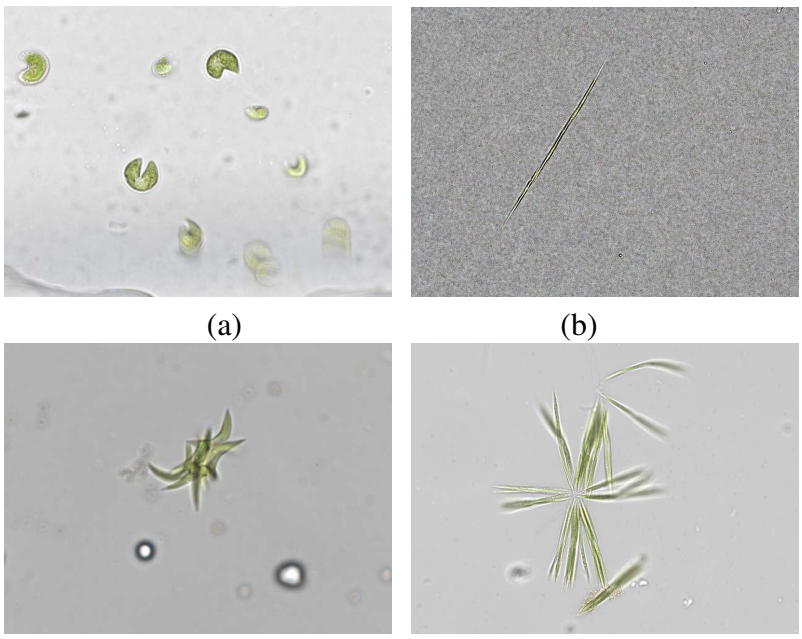
Segmenting green alga images is particularly challenging due to peculiarities resulting both from the application domain and the image acquisition process. Traditional and basic segmentation techniques, such as Canny filter, edge detectors or thresholding, cannot produce accurate segmentations. Studies were thus conducted to analyze appropriate segmentation techniques to these images. First, a contour-based approach based on the level set method has been considered, since it is intuitive to evolve a dynamic curve towards alga boundaries. Due to limitations of the level set method, a region-based segmentation technique derived from the region growing principle has been explored with the application of image enhancement procedures as a preprocessing step.

This chapter is organized as follows: Section 3.1 discusses the challenges for segmenting green microalga images. Section 3.2 outlines a procedure for the automatic sampling of intensities in image regions that is employed in both segmentation methods. Section 3.3 describes the level set method proposed for green alga segmentation, whereas Section 3.4 details the segmentation methodology based on the region growing principle. Section 3.5 presents the experiments conducted for evaluating the proposed segmentation techniques. Section 3.6 summarizes the contributions and results presented in this chapter.

3.1 Problem characteristics

The biologists provided microscope images which confirm that alga species from the Selenastraceae family present very diverse shapes. Such natural diversity added to other peculiar image characteristics resulting from the acquisition process render the automatic segmentation of these images a very difficult task. Thus, a segmentation method shown to be effective on these samples is likely to perform well also on other samples depicting less complex microalga families, providing an essential tool for further developments in the computational support to the task of taxonomical classification.

The data under investigation is a set of 600×800 pixels colored microscope images, with 8 bits per channel. Each image depicts one or multiple alga regions, all from a single species, as it captures an observation of a specific cultured strain under the microscope. Figure 30 depicts some green alga images with a noticeable intensity variation in their background, and a smooth intensity variation within the background region of each image is also observable.



(a) (b)

(c) (d)

Figure 30 – Examples of green microalga images: (a) image characterized by the presence of noise, artifacts and small objects; (b) elongated single alga image characterized by the presence of mucilage in its bottom corner; (c) image depicting colonies where multiple alga cells overlap; (d) a colony of multiple elongated cells.

The images also present noise and artifacts. Figure 30(a) shows an image with some dark particles and an artifact at the bottom. The background of Figure 30(b) is extremely noisy and the alga cells color is not green. The image in Figure 30(c) presents round-shaped algae with bright interiors and dark boundaries. Moreover, although alga cells are usually green, they are brighter near the corners, as shown in Figure 30(d). This is due to alga movement under the microscope lens.

Bearing these scenarios in mind, it is clear that accurate segmentation of green algae must handle the heterogeneous conditions of background regions and identify the alga regions that might be blurred and do not present intensities similar to green. Furthermore, segmentation must be robust to noise and artifacts, extracting only those structures relative to alga cells.

In order to correctly capture the color variations typical of background and alga regions an automated procedure for sampling intensities of image regions is presented next, so that which will be later employed in the proposed segmentation methodologies.

3.2 Automatic sampling procedure

The proposed segmentation methodologies require sampling colors from the regions associated with the alga cells [163]. For that purpose, the paths representing the pixels of such sample regions are computed from a mask. This is a binary image automatically obtained from the original RGB image, providing an initial segmentation that is used to obtain an improved segmentation.

The covariance matrix of the preprocessed RGB image is used to compute the mask, which identifies the target foreground and background regions. The first step relies on computing the local mean values μ_L relative to each point in the image domain, i.e., $\mathbf{x} \in \Omega$:

$$\mu_L(\mathbf{x}) = \frac{1}{|\Omega|} \int_{\Omega} I(\mathbf{x}-\mathbf{y}) d\mathbf{y}, \tag{3.1}$$

$$A(\mathbf{x}) = I(\mathbf{x}) - \mu_L(\mathbf{x}). \tag{3.2}$$

Then a local covariance matrix $C(\mathbf{x})$ of the color channels relative to each domain point is computed, given by:

$$C(\mathbf{x}) = A(\mathbf{x})^T A(\mathbf{x}). \tag{3.3}$$

Finally, the eigenvalues and eigenvectors of the covariance matrix $C(\mathbf{x})$ are computed:

$$V^{-1} C(\mathbf{x}) V = D, \tag{3.4}$$

in which V is the matrix of eigenvectors and D is a diagonal matrix of the eigenvalues of $C(\mathbf{x})$, given by $v = \{D_{1,1}, ..., D_{m,m}\}$. The eigenvalues, which are computed for each pixel, can be represented as m images, each one capturing the image properties from a different perspective. A visual inspection of these eigenvalue images led to the choice of the third eigenvalue image as the most effective to capture the alga characteristics.

The mask B_M that flags pixels as associated with either alga or background regions is obtained by thresholding the third eigenvalue image using its mean intensity value. As a result,

the algae-related pixels are one-valued in B_M, whereas the background pixels are assigned zero values. Finally, a morphological erosion operation is performed in B_M in order to remove false responses that might arise on the background. For this operation, a disk of radius 3 is used as the structuring element.

Figure 31 illustrates the process of obtaining the binary mask B_M departing from the original image depicted in Figure 31(a). The smoothed image obtained from an anisotropic diffusion filtering is depicted in Figure 31(b). Figures 31(c), 31(d) and 31(e) depict the images constructed from the first, second and third eigenvalues of each point in the image domain. After selecting the third eigenvalue image and thresholding it by its mean intensity value, the mask B_M shown in Figure 31(f) is obtained.

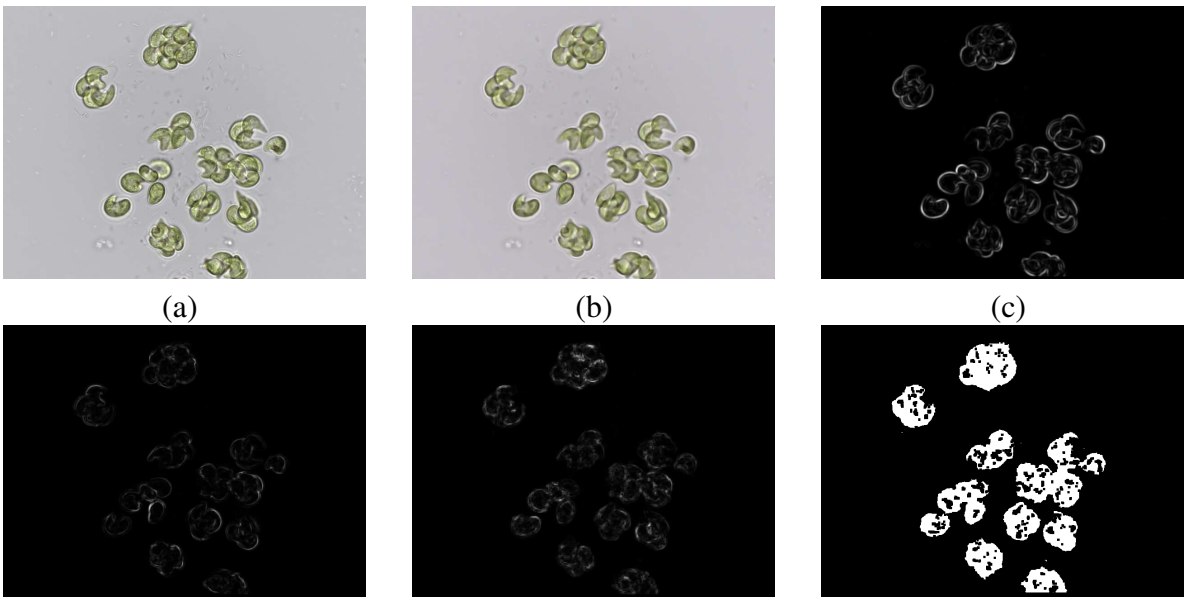


Figure 31 – Illustration of the steps in computing the mask: (a) original image; (b) filtered image using anisotropic diffusion; (c-d-e) the images representing the computed eigenvalues; (f) the mask obtained after thresholding the third eigenvalue image by its mean intensity value.

3.3 Segmentation based on the level set method

3.3.1 Traditional level set

Consider $\Gamma : [0,1] \times [0,\infty) \rightarrow \Omega$ as the parametric curve that divides the domain in foreground (Ω_1) and background ($\Omega \backslash \Omega_1$) regions. In the level set method, the dynamic curve is a *Lipschitz* function $\phi : \Omega \rightarrow \mathbb{R}$, also called level set function, that can be interpreted as the zero-level of a function in higher dimension, for which:

$$\phi(\mathbf{x},t) = \begin{cases} < 0 & \text{if } \mathbf{x} \text{ is on the inside relative to } \Gamma(t) \\ 0 & \text{if } \mathbf{x} \text{ is on } \Gamma(t) \\ > 0 & \text{if } \mathbf{x} \text{ is on the outside relative to } \Gamma(t) \end{cases}$$

Here, $\phi(\mathbf{x},t)$ refers to the curve position in the domain Ω at a given time step t. The level set function ϕ evolves through the domain Ω according to a speed function F, given by the level set equation:

$$\frac{\partial \phi}{\partial t} + F|\nabla \phi| = 0, \tag{3.5}$$

knowing that $\phi(\mathbf{x},0) = \phi_0(\mathbf{x})$ is the initial position of the curve in Ω. The level set function is usually defined by means of a signed distance function, such as the Euclidean distance :

$$\phi(\mathbf{x},t) = \begin{cases} -d_E(\mathbf{x},\Gamma(t)) & \text{if } \mathbf{x} \text{ is on the inside relative to } \Gamma(t) \\ d_E(\mathbf{x},\Gamma(t)) & \text{if } \mathbf{x} \text{ is on the outside relative to } \Gamma(t) \end{cases} \tag{3.6}$$

As for F, the usual choices are the mean curvature [164], in which the speed is defined by the curvature values of curve points, and the geometric term [165] in which an edge potential function is used to stop curve evolution at objects boundaries.

Eq. (3.5) must be solved numerically and the level set function may gradually degrade along successive time steps, due to numerical instabilities. This problem is handled using the reinitializing level set equation, described by:

$$\frac{\partial \phi}{\partial t} = sign(\phi_0)(1 - |\nabla \phi|) \tag{3.7}$$

in which $\phi_0 \approx \phi(x,0)(t=0)$ and sign(ϕ_0) is computed as:

$$sign(\phi(\mathbf{x},t)) = \begin{cases} -1 & \text{if } \phi(\mathbf{x},t) < 0 \\ 0 & \text{if } \phi(\mathbf{x},t) = 0 \\ +1 & \text{if } \phi(\mathbf{x},t) > 0 \end{cases}$$

Although the level set method is suitable for segmenting algae, its standard formulation is highly sensitive to the initial positioning of the curve. Some authors have incorporated region-based terms into the formulation, or have employed sophisticated optimization schemes [166] [167], in order to address this problem.

The *Chan and Vese* method [168] assumes that images can be described by statistically homogeneous regions and aims to minimize the following energy functional:

$$\begin{aligned} F_{CV}(c_1,c_2,\phi) \;=\; & \lambda_1 \int_\Omega (I(\mathbf{x}) - c_1)^2 H(\phi(\mathbf{x}))d\mathbf{x} + \lambda_2 \int_\Omega (I(\mathbf{x}) - c_2)^2 (1 - H(\phi(\mathbf{x})))d\mathbf{x} + \\ & \nu \int_\Omega \delta(\phi(\mathbf{x}))|\nabla \phi(\mathbf{x})|d\mathbf{x} + \rho \int_\Omega H(\phi(\mathbf{x}))d\mathbf{x}, \end{aligned} \tag{3.8}$$

in which ρ, λ_1, λ_2 and ν are positive parameters acting as weights for their respective terms. Constants c_1 and c_2 are statistical representations (mean intensities) of the foreground and back-

ground regions. $\delta(z) = \frac{d}{dz}H(z)$ is the Dirac function. Heaviside functions $H(\phi)$ allow to represent geometrical quantities and properties of the image domain:

$$H(z) = \begin{cases} 1 & \text{if } z \geq 0 \\ 0 & \text{if } z < 0 \end{cases} \tag{3.9}$$

However, some of the underlying assumptions do not apply to green alga images, which present considerable color variation even within the alga cells. Thus, the proposed method overcomes this limitation by capturing such alga patterns by computing Gaussian distributions that characterize the image regions and incorporating such information into the level set method. This strategy, described in detail in the following section, has been inspired by *Rousson and Deriche*'s method [169].

3.3.2 Proposed level set formulation

Rousson and Deriche (RD) [169] formulated an energy functional based on the level set method which incorporates probability distributions for statistically describing the image regions [170] [171]. For that purpose, it is assumed that the intensities at each point $\mathbf{x} \in \Omega$ are independent and identically distributed by the same random process and the image regions are statistically independent.

Let $P_1(I(\mathbf{x})|\theta_1)$ and $P_2(I(\mathbf{x})|\theta_2)$ be the probability distributions of the foreground and background regions, respectively. Taking the same level set formulation adopted by *Chan-Vese*, but employing the probability distributions to model image regions, *Rousson and Deriche* proposed minimizing the following energy functional:

$$\min_{\phi,\{\theta_1,\theta_2\}} \Bigg\{ F_{RD}(\phi,\{\theta_1,\theta_2\}) = \int_\Omega |\nabla H(\phi)| - \lambda \int_\Omega H(\phi(\mathbf{x})) \log(P_1(I(\mathbf{x})|\theta_1))d\mathbf{x} \\ -\lambda \int_\Omega (1 - H(\phi(\mathbf{x}))) \log(P_2(I(\mathbf{x})|\theta_2))d\mathbf{x} \Bigg\}, \tag{3.10}$$

in which the region parameters $\{\theta_1, \theta_2\}$ are estimated according to an optimization scheme further described elsewhere [169]. Although their method performs well on the green alga images, its formulation does not incorporate edge information, which provides important patterns for characterizing distinct alga shapes.

Motivated by existing strategies [165] [171], an edge potential function has been incorporated into the first term of RD's energy functional, in order to reduce the diffusion process on region boundaries. The underlying rationale is to preserve region edges as much as possible in order to favor accurate identification of alga shapes. Finally, the energy functional of the proposed method is given by:

$$\min_{\phi}\left\{F_{PM}(\phi,\{\theta_1,\theta_2\}) = \int_{\Omega} g|\nabla H(\phi)| - \lambda\int_{\Omega} H(\phi(\mathbf{x}))\log(P_1(I(\mathbf{x})|\theta_1))d\mathbf{x} \right.$$
$$\left. -\lambda\int_{\Omega}(1-H(\phi(\mathbf{x})))\log(P_2(I(\mathbf{x})|\theta_2))d\mathbf{x}\right\}. \tag{3.11}$$

The first term in the functional given by Eq. (3.11) is the length of the contour Γ and the two remaining terms refer to the cost of assigning each domain point inside and outside the contour. The minimization problem defined by Eq. (3.11) is solved by deriving and solving the Euler-Lagrange equations using a gradient descent scheme in relation to ϕ:

$$\frac{\partial\phi}{\partial t} = div\left(g\frac{\nabla\phi}{|\nabla\phi|}\right) + \lambda\log(P_2/P_1) \tag{3.12}$$

in which g is given by Eq. (2.2), defined in Section 3.4.1. The dynamic curve evolves according to the log-likelihood test defined by the second term in Eq. (3.12). If:

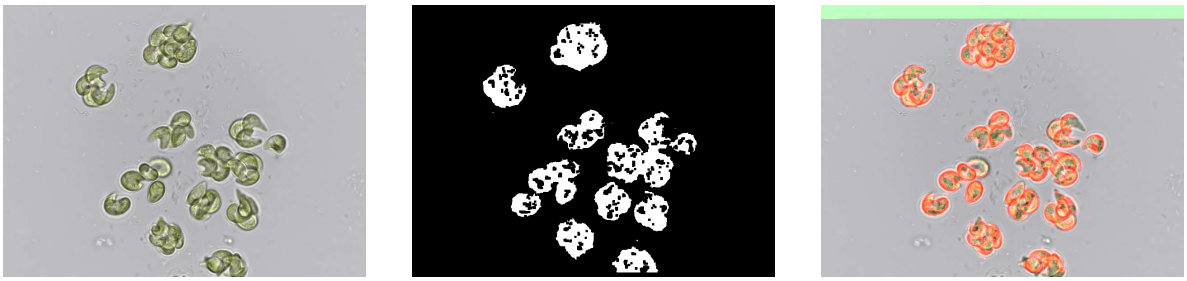
$$\log\left(\frac{P_2(I(\mathbf{x})|\theta_2)}{P_1(I(\mathbf{x})|\theta_1)}\right) < 0, \tag{3.13}$$

the pixel $I(\mathbf{x})$ is likely to belong to the green alga region, otherwise such test indicates that $I(\mathbf{x})$ is likely to belong to the background.

Multivariate Gaussians have been adopted to describe the image regions due to their effective capability in approximating the image's histogram, but considering the intensity variation within image regions, and the straightforward estimation of the associated parameters [172]. The distribution parameters $\theta_1 = \{\mu_1,\Sigma_1\}$ and $\theta_2 = \{\mu_2,\Sigma_2\}$ are the mean and the covariance matrix of the alga and the background regions, respectively. The associated Gaussian distributions are computed as:

$$P_i(I(\mathbf{x})|\{\mu_i,\Sigma_i\}) = \frac{1}{\sqrt{(2\pi)^m|\Sigma_i|}}\exp\Big(-\frac{1}{2}(I(\mathbf{x})-\mu_i)^T\Sigma_i^{-1}(I(\mathbf{x})-\mu_i)\Big) \tag{3.14}$$

in which values for parameters θ_1 and θ_2 are estimated once prior to the curve evolution, and kept fixed during the optimization process of Eq. (3.11). Parameter estimation consists of sampling pixel intensities of both image regions using the mask image obtained from the process described in Section 3.2. Figure 32 illustrates the sampling steps, in which Figure 32(a) shows the original green alga image and Figure 32(b) depicts the mask. Figure 32(c) illustrates the obtained patches for sampling intensities from the target regions of interest, the algae (shown in red) and the background (in green). For performance reasons, as the most image pixels belong to the background, it is sufficient to sample only 10% of them.



(a) (b) (c)

Figure 32 – Examples illustrating the image sampling procedure: (a) original RGB image; (b) generated mask; (c) mask used for sampling, in which the red patches are related to alga pixels and the green patch with the background.

The binary image obtained might include alga cells presenting small holes, because some pixels inside alga regions are more similar to the pattern intensities of the background. This issue can be minimized by applying a dilation morphological operation over the binary image using as structuring element a disk of radius 3. The goal is to fill such holes in the alga regions and to smooth their boundaries, which may also present small concavities. Furthermore, small regions that are not related to alga cells are removed by ignoring regions with perimeters smaller than 50 pixels.

In summary, the proposed level set methodology for segmenting green alga images consists of these major steps:

1. Compute the mask and sample pixel intensities from both the alga and the background regions.

2. Estimate the Gaussian probability distributions parameters $\theta_1 = \{\mu_1, \Sigma_1\}$ and $\theta_2 = \{\mu_2, \Sigma_2\}$, given the alga and background region samples, respectively.

3. Compute the multivariate Gaussian distributions P_1 and P_2 for those regions, according to Eq. (3.14).

4. Initialize $\phi_0 \approx \phi(\mathbf{x}, 0)$ using Eq. (3.6).

5. Use a finite difference approach [80] to compute the numerical solution of Eq. (3.12) in relation to the level set function ϕ. For this purpose, assume that $\Omega \subset \mathbb{R}^2$ and $\mathbf{x} = (x, y)$, so the level set function and the digital images are discretized as bidimensional $[m_1, m_2]$ matrices, in which $\phi(\mathbf{x}) \approx \phi_{i,j}$ for $i = 1, ..., m_1$ and $j = 1, ..., m_2$. The level set curve evolution stops once it reachs the alga boundaries.

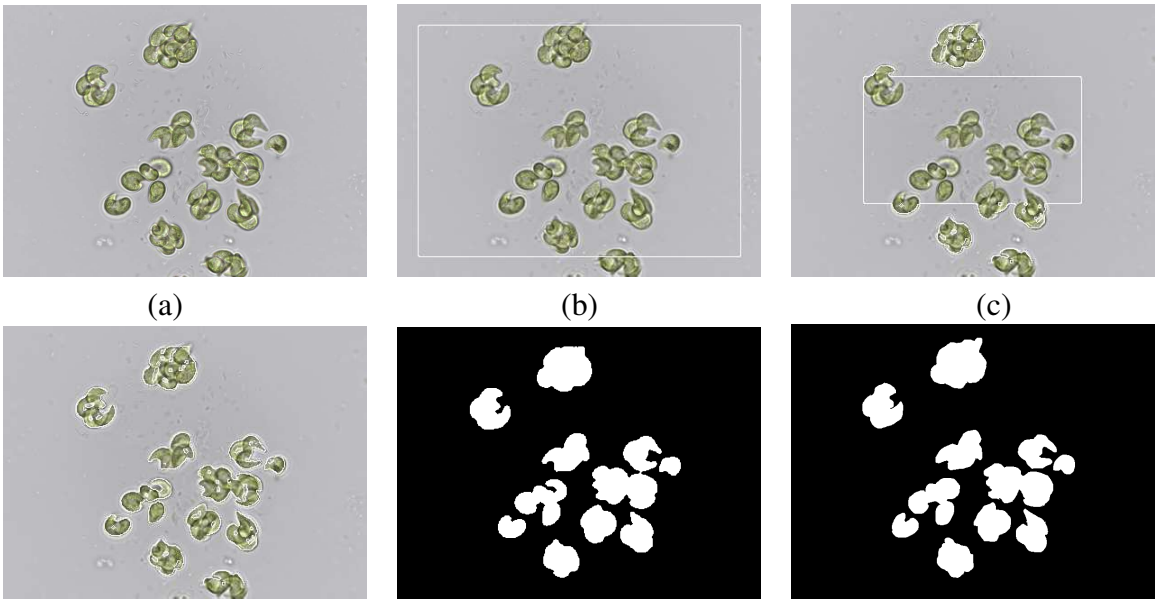
6. Once ϕ converges, generate the binary image by thresholding the final level set as $\phi(\mathbf{x}) < 0$.

7. In the binary image, keep the regions with perimeters greater than 50 (measured in units defined by the uniform spacing of the underlying implicit image grid) and perform a

dilation morphological operation using a disk structuring element of radius 3 to smooth the alga boundaries.

The value 50 was set for the structuring element in the dilation operation since it is enough to remove small groups of pixels which are not related to alga regions. Moreover, such value guarantees that smaller algae are kept in the segmentation.

Figures 33, 34 and 35 illustrate three segmentation cases using the proposed level set approach. The sequence of Figures (a), (b), (c-d), (e) and (f) denote, respectively, the original green alga image, the initial positioning of the level set function $\phi(\mathbf{x},0)$, the two intermediate stages of ϕ, the final positioning of ϕ after the convergence, the binary image depicting the segmentation result after thresholding ϕ, and the ground-truth image. The number of iterations until the level set reaches the final positioning is indicated in the figure captions.



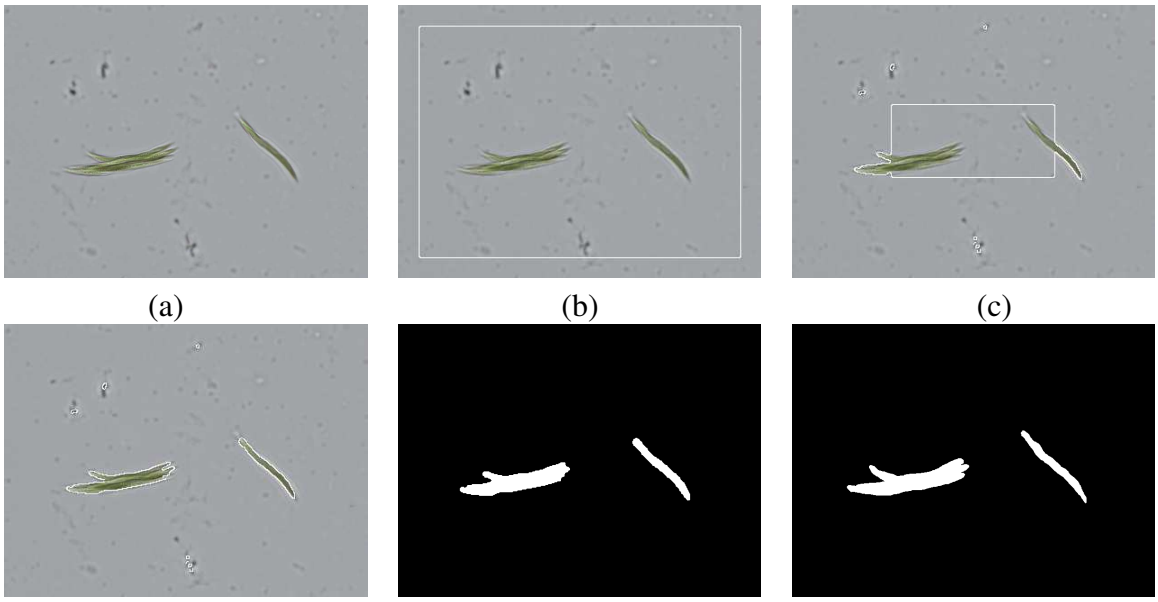
(a) (b) (c)

(d) (e) (f)

Figure 33 – Segmentation of several alga colonies of species *Kirchneriella aperta*: (a) original RGB image; (b) initial $\phi(\mathbf{x},0)$; (c-d) intermediate states of ϕ after 810 iterations; (d) final level set function after $1,780$ iterations; (e) segmentation result after thresholding ϕ and performing a dilation operation; (f) ground-truth image.

Figure 36 illustrates two segmentation cases with unsatisfactory results, in which the top images are the original and the bottom images show the corresponding segmentations. Figures 36(a) and 36(c) present a result in which an area associated to a second alga body has been detected, but its size is sufficiently large not to be removed by the post-processing step. The poor accuracy in segmentation shown in Figures 36(d) is due to transparencies in the alga body and its corners.

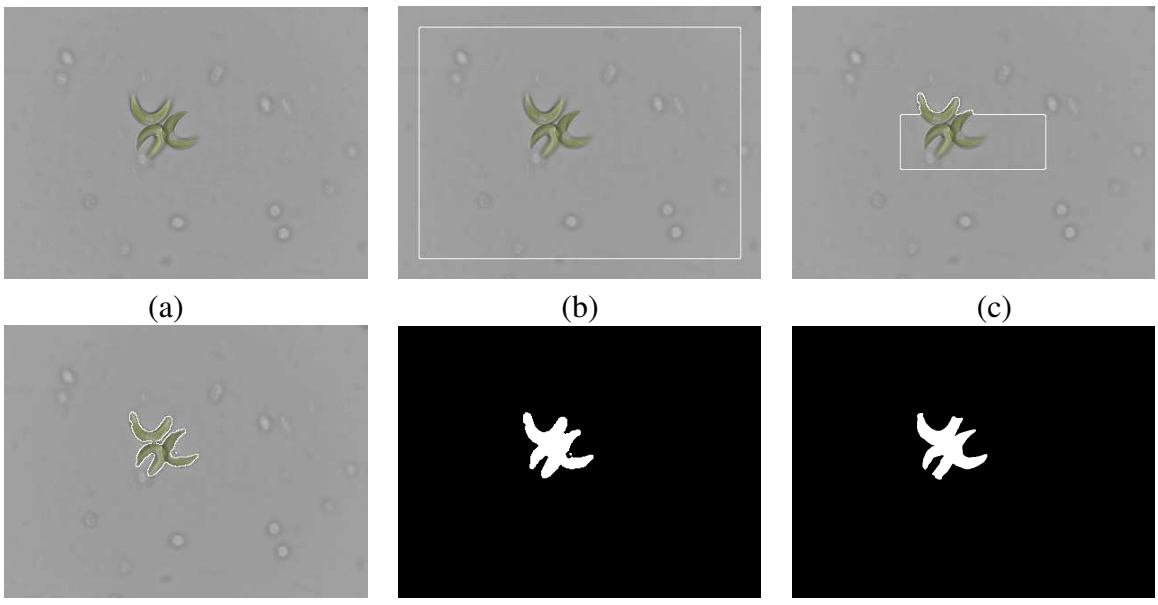
The proposed level set method is robust to initial conditions, i.e., the initial placement of the level set curve does not affect the final segmentation result. The images in Figure 37 show two test cases considering distinct initial placements of the level set curve. Figure 37(c)



(a) (b) (c)

(d) (e) (f)

Figure 34 – Segmentation of two alga colonies of species *Ankistrodesmus densus*: (a) original RGB image; (b) initial $\phi(\mathbf{x},0)$; (c) intermediate states of ϕ after $1,110$ iterations; (d) final level set function after $1,600$ iterations; (e) segmentation result after thresholding ϕ; (f) ground-truth image.



(a) (b) (c)

(d) (e) (f)

Figure 35 – Segmentation of an alga colony of species *Selenastrum bibraianum*: (a) original RGB image; (b) initial $\phi(\mathbf{x},0)$; (c) intermediate states of ϕ; (d) final level set function after $1,710$ iterations; (e) segmentation result after thresholding ϕ; (f) ground-truth image.

shows the two initial positions of the level set ϕ. Figures 37(d) shows intermediate states and Figure 37(e) illustrates the final positions after $1,620$ iterations. Figure 37(f) present the resulting alga segmentations. As the proposed method does not optimize the parameter distributions while minimizing the energy function of Eq. (3.11), it is less sensitive to initial conditions as compared with the original RD's model.

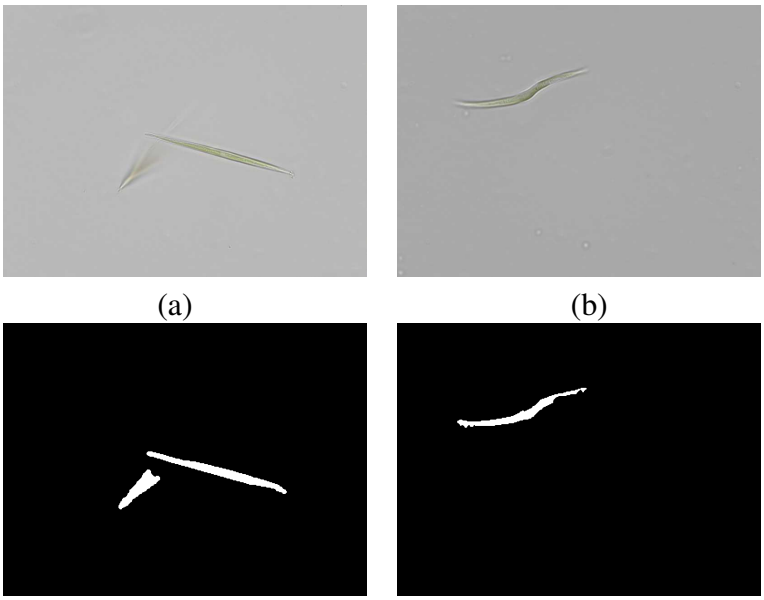


Figure 36 – Low accuracy segmentations obtained with the proposed method: (a-b) original images; (c-d) corresponding segmentations.

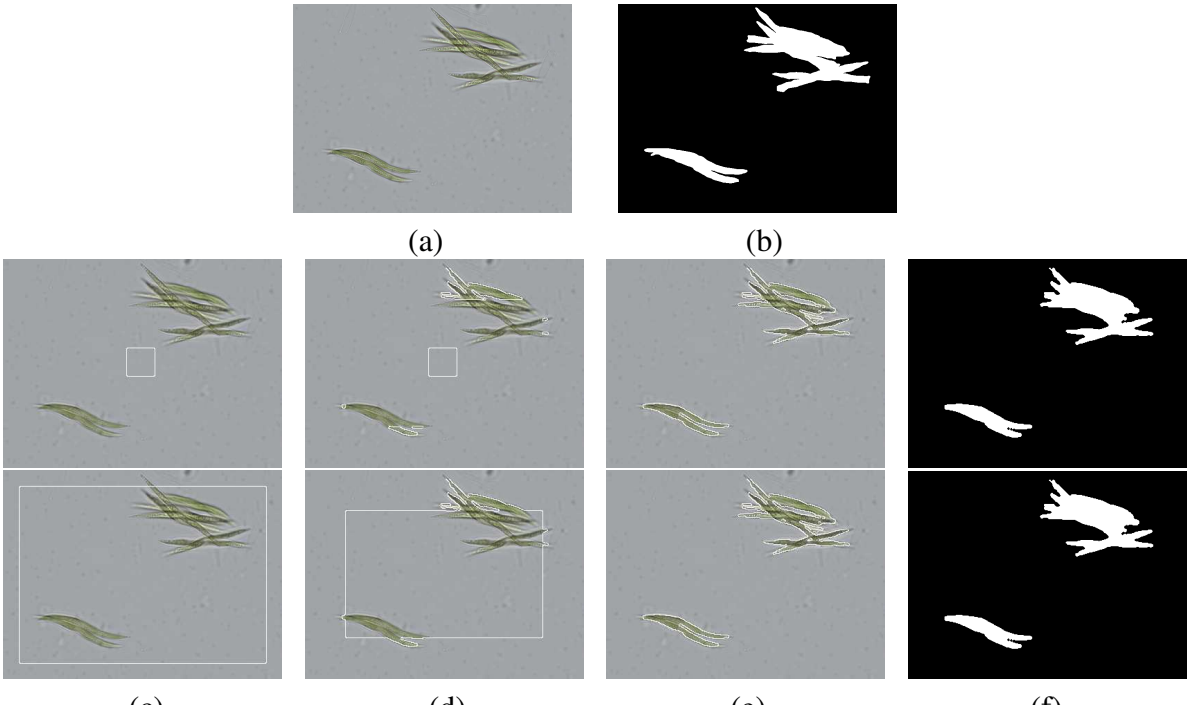


Figure 37 – Level set evolution from two distinct initial positions (at each row): (a) original RGB image; (b) ground-truth image; (c) initial $\phi(\mathbf{x},0)$; (d-e) intermediate state of ϕ; (f) after convergence of ϕ. [13]

A visual inspection of the generated segmentations allows to infer that the proposed method satisfactorily identified the green alga shapes. It is noticeable that the level set curve correctly surrounded the alga cells as a result of the appropriate representation of image regions as Gaussian distributions, which captures the intensity patterns according to the sampling procedure. The contour-based approach is robust: it successfully handled intensity variations in the

background and disregarded any unwanted objects appearing in the background with intensity patterns dissimilar to algae, whereas small objects of color similar to green algae detected in the segmentation are later removed in the post-processing step.

In general, it was observed that the multivariate Gaussians are effective statistical representations of the image regions. Such models can capture the patterns and details characteristic of each region, and the sampling procedure ensures that distinguishing distributions are computed between them, yielding to a faster evolution of the dynamic curve towards the desired image regions. Unlike *Rousson and Deriche*'s model, the distribution parameters in the proposed method are computed *a priori* and remain fixed along the level set function optimization, leading to faster convergence and less sensitivity to initial conditions.

Also, the method's energy functional is non-convex due to the optimization space of Heaviside functions, meaning that infinite solutions are valid representations of a given optimal Ω_1. The practical implication is that the method is weakly sensitive to initial conditions and local optimal solutions are obtained by computing numerical solutions to the Euler-Lagrange equations associated with gradient descent schemes.

As further detailed in the experimental results (Section 3.5), the proposed method is computationally expensive, since it depends on the convergence of the numerical equations. Another weaknesses of the method are the preservation of certain peculiar structures that appear in some specific alga species such as mucilage, and its failing to group some colony cells with transparent areas, specifically in regions where alga cells join. Such practical limitations led us to devise a more efficient and precise segmentation approach.

3.4 Segmentation based on region growing

This section describes a faster and more precise segmentation methodology for green alga images based on region growing when compared to the previous method. Images are initially preprocessed for noise suppression and then transformed to the Hue-Saturation-Value (HSV) space in order to reduce the color variation in their regions. A contrast enhancement in the hue channel is then performed using an equalized version of the Value channel. The mask enables to define the proper number of seed points, avoiding undesirable situations of missing relevant regions or placing multiple seeds in a single region. Moreover, that image is also used to sampling intensities of the alga and background regions in the enhanced Hue channel in order to estimate their associated Gaussian distributions. The region homogeneity criterion to guide region growth is set by performing likelihood tests on the estimated Gaussian distributions. Finally, alga regions are smoothed with a morphological operation based on the rolling ball operator [173].

3.4.1 Preprocessing steps

Some preprocessing steps are executed to improve image quality before applying the region growing algorithm. The first step smooths the original RGB image prior to obtaining its corresponding HSV representation. Further processing is applied to the Hue channel of the HSV representation, which provides sufficient contrast to distinguish between the alga and the background regions. A contrast enhancement is then applied to generate a smooth image with highlighted alga regions.

HSV Model

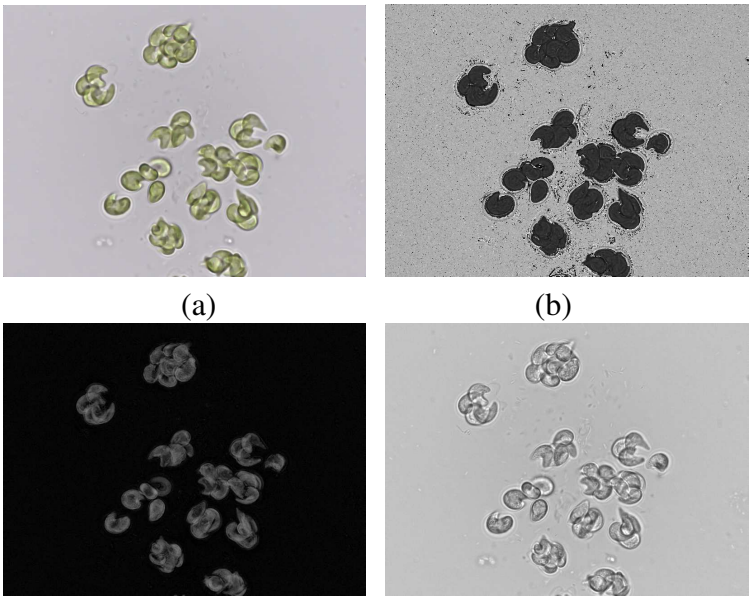


Figure 38 – Green alga image transformed to the HSV representation: (a) filtered RGB image; (b) Hue channel; (c) Saturation channel; (d) Value channel.

In this method, the HSV representation of the filtered RGB image I_{ADF} has been chosen since it is more effective in capturing the contrast between the alga cells and the background. The corresponding HSV model of a given RGB input image may be computed with the equations given in [174]. Figure 38 illustrates the results of the conversion for a particular green alga image. Figure 38(a) presents the original RGB image, whereas Figures 38 (b), (c) and (d) show the Hue (H), Saturation (S) and Value (V) channels of the corresponding HSV image. It is noticeable in the image showing the Hue channel, depicted in Figure 38(b), that alga cells are characterized by a uniform gray intensity, whilst the background is noisy – it is possible to observe the lighting variation in the background areas. The Saturation channel, shown in Figure 38(c), presents the alga regions in brighter intensities with blurred boundaries and it is discarded in the subsequent steps. The Value channel image depicted in Figure 38(d) is simply a grayscale image, and is used in the following step responsible for the Hue contrast enhancement.

Image enhancement

A contrast enhancement procedure is applied to the Hue channel in order to enable a more accurate identification of the alga regions. The procedure considers the intensity information registered in the Value channel, since the Hue channel alone may not disclose sufficient information. The rationale is to perform a histogram equalization in the Value channel that makes possible the analysis of the intensity variation of the background pixels and then identify those intensity levels most likely associated with alga pixels. Such intensities are determined by thresholding the equalized image, thus generating a binary image that is used to weight and highlight the alga pixels in the Hue channel.

The histogram equalization generates a new image by quantizing the intensities in the Value channel to a predefined number of discrete gray levels. The pixel values are roughly uniformly distributed across the quantized gray level bins, so that the resulting histogram is approximately flat. After a visual observation of the intensities of green alga regions for several kinds of equalized images, the Value channel is transformed to 64 intensities for better discrimination of algae-related pixels, obtaining the equalized image I_{EQ}.

Algae regions are typically associated with the lower intensities in the histogram of image I_{EQ}. As the idea is to obtain a binary image that flags the algae-related pixels for enhancement, the threshold value τ is determined by considering the associated intensities in the histogram and the perimeter of the candidate alga regions in the mask. A histogram analysis revealed that the histogram of I_{EQ} has non-zero values at levels 1, 5, 9 and 12. However, gray level frequencies are highly dependent on the number of pixels that belong to alga regions. Thus, the equalized image I_{EQ} is binarized accordingly using the larger perimeter p from the candidate alga regions in the mask. The value of τ is determined as:

$$\tau = \begin{cases} 1, & \text{if } p \leq 500 \\ 5, & \text{if } p > 500 \text{ and } p \leq 1000 \\ 9, & \text{if } p > 1000 \text{ and } p \leq 1500 \\ 12, & \text{otherwise.} \end{cases} \tag{3.15}$$

Once τ is obtained, the binary image B_{EQ} is computed as:

$$B_{EQ}(\mathbf{x}) = \begin{cases} 1, & \text{if } I_{EQ}(\mathbf{x}) \geq \tau \\ 0, & \text{otherwise,} \end{cases} \tag{3.16}$$

where B_{EQ} is the binary mask image flagging the algae-related pixels. The image enhancement will weight such pixels to emphasize their intensities, while preserving background patterns. This operation takes into account the perimeter p, since for images containing small or thin alga the Hue channel typically has low contrast due to an unbalanced amount of background pixels.

Eq. (3.17) describes the enhancement process for alga images with $p > 250$, which doubles the intensities of the background pixels in the Hue channel I_{hue}:

$$I_H(\mathbf{x}) = \begin{cases} 2I_{hue}, & \text{if } B_{EQ}(\mathbf{x}) = 1 \\ I_{hue}, & \text{otherwise.} \end{cases} \tag{3.17}$$

When $p < 250$, $I_H = I_{hue} + B_{EQ}$ once the Hue channel does not present a high contrast between alga and background regions.

The intermediate image I_H in the above equation refers to the updated Hue channel with the alga regions highlighted. Finally, I_H is normalized to $[0,1]$ resulting in the final enhanced image I_{EN}, which displays a better visual contrast between the alga and the background regions, as compared to the original Hue channel. The steps of the enhancement procedure are illustrated in Figure 39. Figure 39(a) shows the original Hue channel image. Figure 39(b) presents I_{EQ}, the Value channel image after histogram equalization, in which pixels associated with alga cells have the lowest intensities. Figure 39(c) shows the weighted image B_{EQ} that indicates which pixels must be enhanced in the Hue channel. Finally, Figure 39(d) depicts the final enhanced Hue channel image, in which alga regions are noticeably emphasized and the intensity patterns have been preserved.

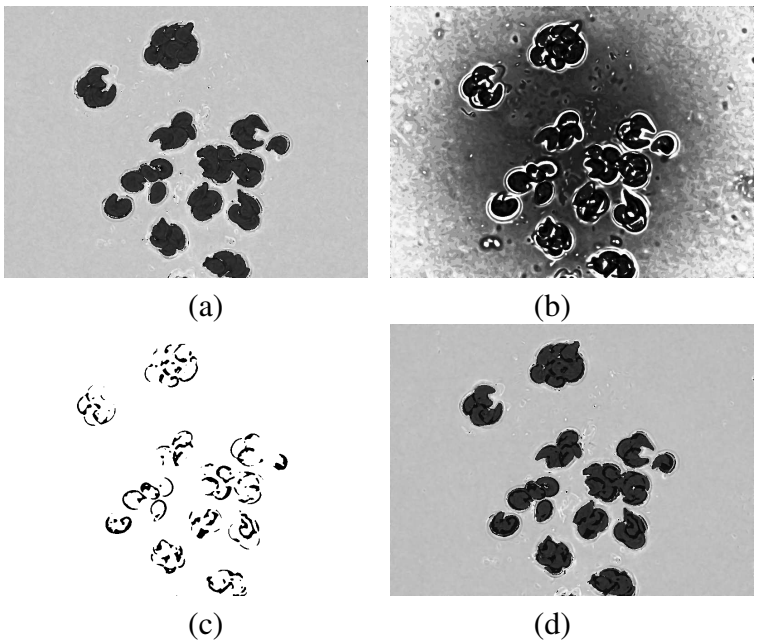


Figure 39 – Image enhancement: (a) the Hue channel image I_{hue}; (b) the equalized Value channel image I_{EQ}; (c) the binary image B_{EQ}; (d) the enhanced Hue channel image I_{EN}.

Once the preprocessing steps are finished, the appropriate Hue channel I_{EN} is obtained for the upcoming segmentation process, in which alga cells have distinct intensities with respect to the background pixels.

3.4.2 Seeded region growing

The seeded region growing algorithm, or simply region growing algorithm, operates by grouping (i.e., growing) pixels or subregions into larger regions based on a predefined similarity criterion [174]. Some seed pixels are initially selected based on some criterion (e.g. color, intensity, or texture). Once the initial seeds are placed in the image domain, the growth process seeks to obtain homogeneous image regions, i.e., it tries to find an accurate segmentation of the image into regions with the property that each connected component of a region contains exactly one of the initial seeds. The presence of noise may result in oversegmentation, which is typically handled with a subsequent region merging process.

Two major concerns must be handled when performing a segmentation with region growing: where to place the initial seeds in the image domain and which homogeneity criterion should be adopted to characterize the image regions. As for the seed placement problem, it is expected that segmentation of an image composed by N relevant target objects should start with N initial seeds, one located at each object. As for the region growing, the homogeneity criterion must capture the properties of the target objects.

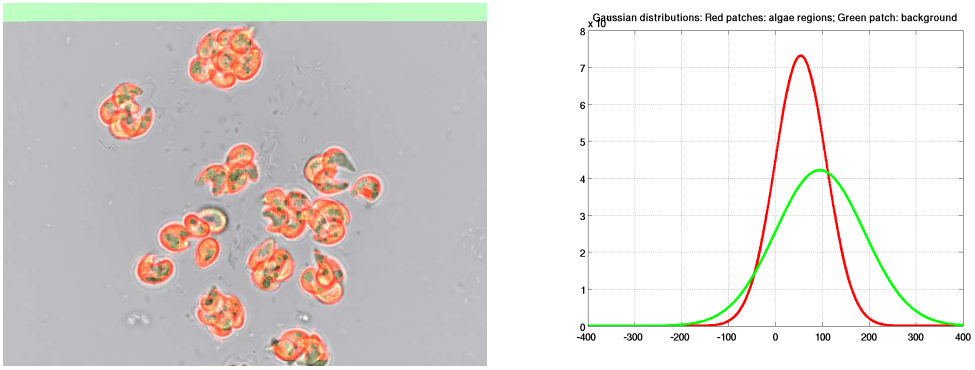
For this specific segmentation problem, each relevant alga region (either a single cell or a colony) would require a seed representative. Thus, an approach to automatically determine where to place the seed points was devised, guaranteeing that one single seed will be placed in the interior of each alga region. The seed placement relies on a mask obtained from the filtered RGB image I_{ADF}, which provides a binary mask useful to determine the seed points.

The homogeneity criterion and the conditional test to drive the region growth must account for the intensity variations of the alga pixels in the Hue channel. The image regions (alga cells and background) are characterized by the Gaussian distributions of their intensities, described by their mean and standard deviation. These parameters are computed by automatically sampling a sub-set of pixels from each region. Any alga regions can be modeled with a single probability distribution, as the pixels associated with algae have the lowest intensities (darker regions) in the Hue channel.

Region sampling

The foreground and background regions in the image may be characterized by their intensity probability model distributions. The parameters characterizing the respective distributions may be estimated by sampling the foreground and background regions as identified in the mask B_M in the enhanced Hue image. The quality of the probability distribution estimation depends on an effective sampling procedure.

The sampling procedure is exemplified in Figure 40. Figure 40(a) illustrates the patches used to sample the target regions, namely the algae (shown in red) and the background (in green). For performance reasons, it is sufficient to sample only 10% of the background pixels.



(a) (b)

Figure 40 – Illustration of the region sampling procedure: (a) the red patches depict the alga region, while the green patch refers to the background region; (b) the estimated Gaussian distributions of the intensities in the alga (red line) and in the background (green line) regions.

Figure 40(b) presents the estimated Gaussian distributions estimated for algae and background, given by their mean and standard deviations as computed from the sampled intensities in their respective regions.

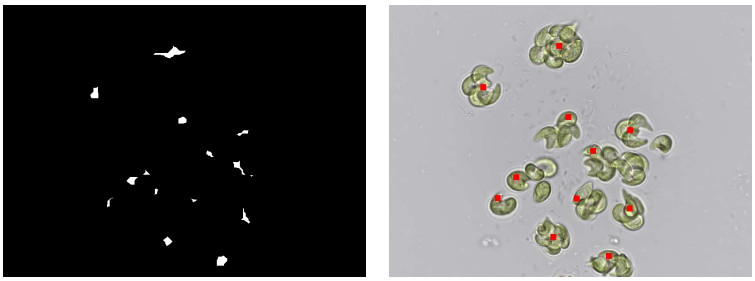
Setting the seeds

The mask possibly includes multiple alga regions, and it is necessary to determine manually a single pixel in each representative alga region. Artifacts characterized by small areas with less than 150 pixels may be found in these images, which must be disregarded. The principle is thus to consider individually each region (with area greater than 150 pixels) of B_M and select a contour point from an eroded version of B_M, thus making sure that the seed points are placed inside the target shapes.

Computing the seed points thus requires the following steps:

1. In an iterative process, perform successive morphological erosion operations on image B_M using a structuring element of size 1 until all regions completely shrink. Let n_I be the number of iterations performed.
2. Erode image B_M using a structuring element of size $\frac{n_I}{2}$, producing a new image B_{Er}.
3. Discard any regions in B_{Er} with perimeter smaller than 150 pixels.
4. Pick the top-leftmost pixel from the external contour of each region i in B_{Gr} as the respective region seed $(s_{i,1}, s_{i,2})$.

The above methodology guarantees that one seed is placed inside each alga region, in which associated region in B_{Er} has a perimeter greater than 150. The size of the structuring element was manually set by taking into account the image resolution and the alga area sizes. This process generates a set of seeds $\mathscr{S} = \{(s_{1,1}, s_{1,2}); \ldots; (s_{N,1}, s_{N,2})\}$ placed in the image domain for the region growing process.



(a) (b)

Figure 41 – Determining the seeds: (a) the image B_{Er}; (b) the seed point placed over the alga regions (shown in red).

Figure 41 illustrates the strategy for computing the seed points. Figure 41(a) shows the image B_{Er} resulting from executing Step 2 of the seed placement method. Figure 41(b) shows the seed point computed (shown in red) placed over the enhanced Hue channel image I_{EN}.

Homogeneity criterion

The homogeneity criterion defines whether a candidate pixel should be incorporated into a specific region. As such, it must consider the statistically relevant patterns of the different image regions, such as color, texture or intensity. Thus, a criterion must be chosen that captures the intensity patterns of the alga and the background regions.

In this case, the Gaussian distributions of the pixel intensities are effective to characterize alga and background regions in the Hue channel. First, the distributions parameters $\theta_1 = \{\mu_1, \sigma_1\}$ and $\theta_2 = \{\mu_2, \sigma_2\}$ are estimated, taking as parameters the means μ_i and the standard deviations σ_i computed from the alga and the background region samples, respectively, in the enhanced Hue image I_{EN}. Distributions P_1 and P_2 are thus computed as:

$$P_i(I(\mathbf{x})|\{\mu_i, \sigma_i\}) = \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left(-\frac{||I(\mathbf{x}) - \mu_i||^2}{2\sigma_i^2}\right) \tag{3.18}$$

in which $||.||$ refers to the Euclidean norm. The distinctive Gaussian distributions of both regions are clearly depicted in the corresponding plots shown in Figure 40(b).

Region growth process

This process can be interpreted as a pixel labeling procedure in which all pixels belonging to a homogeneous region will be assigned the same label. The seed pixel is compared with its neighboring pixels, and they are grouped into a single region if the homogeneity criterion is satisfied. The region growing finishes once all pixels have been assigned a region label, which does not require a merging process.

Let $\mathscr{S}$ be the set of seed points, in which $s_i \in \Omega$, and let P_1 (P_2) be the probability distribution associated with all the alga regions (background). Considering an 8-connectivity

neighborhood, each image pixel is tested to verify whether it satisfies the criterion for inclusion in an alga region. If:

$$\left.\frac{P_2(I_{EN}(\mathbf{x})|\{\mu_2,\sigma_2\})}{P_1(I_{EN}(\mathbf{x})|\{\mu_1,\sigma_1\})}\right) < 0, \tag{3.19}$$

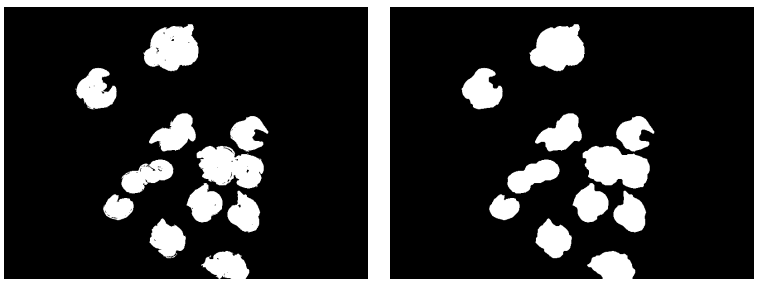
x belongs to an alga region and **x** belongs to the background otherwise.

A binary image is obtained, where pixels that satisfy the conditional (domain points likely to belong to an alga region) are assigned a value 1, otherwise pixels are assigned a value 0. The next step relies on appending to each seed point in $\mathscr{S}$ all the one-valued points in the binary image which are 8-connected to it, resulting in an image with connected components corresponding to each alga cell, colony and background areas. In this process, the probabilities of each neighbor pixel **y** and the region associated with a seed s are computed as follows:

$$d(\mathbf{y}) = |P_1(I_{EN}(\mathbf{y})|\{\mu_1,\sigma_1\}) - P_1(I_{EN}(s)|\{\mu_1,\sigma_1\})|. \tag{3.20}$$

Then **y** is merged into the region associated to a seed s when condition $d(\mathbf{y}) < D_0$ is satisfied. D_0 is the average value of $d(\mathbf{y})$, the one-valued pixels in the mask B_M. Finally, each connected component receives a distinct region label, uniquely identifying each alga region. Additionally, a binary image B_{RG} is generated by keeping only the alga regions (as one-valued intensities) associated with seed points.

Figure 42 illustrates the region growing process using the enhanced Hue channel image I_{EN} as input and the seed points indicated by the red markers in Figure 41(b). The result is shown in Figure 42(a), in which the white areas correspond to pixels that are similar to the respective alga regions in terms of hue intensity.



(a) (b)

Figure 42 – Post-segmentation: (a) result of the region growing process for a particular image; (b) binary image representing the final segmentation, after applying the rolling ball.

3.4.3 Rolling ball transformation

Algae movement during image acquisition leads to blurred corners and/or some transparent parts in the alga cells. As a result, in some cases the segmented shapes in B_{RG} might present small concavities and holes, as observed in the image depicted in Figures 42(a) and 43(a). This

problem is handled by applying a rolling ball transformation to fill in any undesirable holes or concavities in the shapes obtained with the region growing process. The rolling ball operation produces another binary image I_{SEG} that denotes the final segmentation.

(a) (b)

Figure 43 – Rolling ball transformation: (a) binary image resulting from the region growth process; (b) result obtained by the rolling ball.

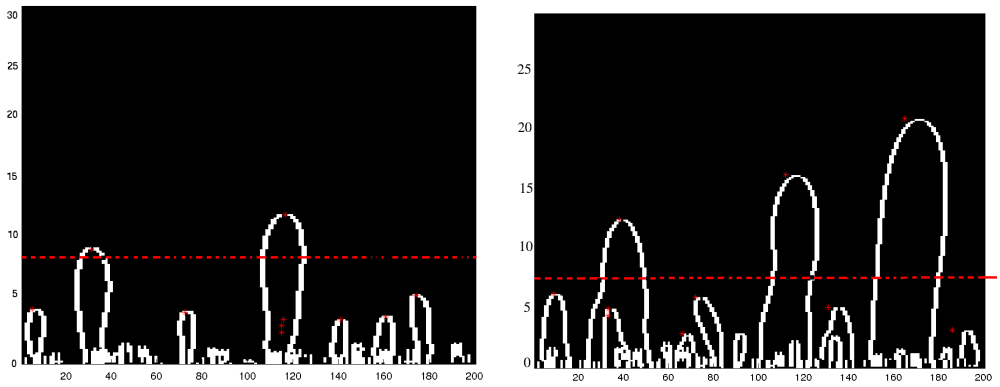
The rolling ball transformation [175] can be described as a morphological closing of the target region, followed by a hole filling operation. A hole is a set of background pixels that cannot be reached by filling in the background from the edge of the image. In the rolling-ball transformation, a disk structuring element with a predefined radius is applied to the binary images.

Determining the size of the disk radius is difficult, since the alga shapes differ a lot in size and complexity. Setting a single radius length to handle all shapes would likely result in poor quality segmentation in some cases. Radius size is automatically computed in each case, taking into account the perimeter of the alga contours and some prior knowledge on their shapes. First, small algae do not require large disks for the rolling ball, because it is important to preserve their characteristic concavities. On the other hand, for alga colonies it is better to adopt medium-sized radii, since junctions between cells must be preserved. Finally, larger radii can be used on elongated alga shapes, which often present blurred corners and the rolling ball operation will not affect the shape essence in this case.

Shape complexity is determined by identifying the number of peaks in its corresponding *Curvature Scale Space* (CSS) map [176]. The CSS descriptor captures the key local shape features by representing the shape boundary curvatures in a scale space which describes the locations of convex (or concave) segments and also detects the degree of convexity (or concavity) of such segments. The scale space representation of a shape is created by tracking the position of inflection points in a shape boundary filtered by low-pass Gaussian filters of variable widths. As the width of the Gaussian filter increases, negligible inflections are removed from the boundary and the shape becomes smoother. The remaining inflection points in the representation are likely to describe relevant object characteristics.

The result of this multi-scale smoothing process is a map depicting an interval tree formed by several inflection points. The shape contours have been subsampled to 200 points. Figures 44(a) and 44(b) show the CSS maps of one alga shape from those depicted in Figures 42(a) and 43(a), respectively, in which the red points are the maxima. The $x-$axis presents

the arc length of the alga contour after subsampling to 200 points. The $y-$axis refers to the width of the Gaussian low-pass filtering in the contour. It is noticeable that maps of colonies have more points of maximum than maps of single alga.



(a) (b)

Figure 44 – Curvature Scale Space maps: (a) map of the alga colony shape depicted in Figure 42; (b) map of the single alga shape depicted in Figure 43(b).

The number of peaks is identified by thresholding the CSS map at the Gaussian width 7, indicated in Figure 62 by the dashed red lines. Defining N_{CSS} as the number of peaks and p the shape perimeter, the disk radius is computed as:

1. **IF** $p > 250$ **AND** $N_{CSS} \leq 3$
2. **THEN** radius $\leftarrow 6$;
3. **ELSE** radius $\leftarrow 2$.
4. **END IF**

Figure 43(b) shows the result of applying the rolling ball operator to the binary image in Figure 43(a), in which holes and concavities were filled. Figure 42(b) presents the final segmentation result after the rolling ball transform produced smoother alga regions when comparing with the output image of the region growing, shown in Figure 42(a).

The proposed segmentation methodology based on region growing can be summarized by the following steps:

- **Preprocessing**

  1. The original RGB image I is filtered with the anisotropic diffusion filter, producing a filtered image I_{ADF};
  2. The RGB image I_{ADF} is converted into its HSV representation, yielding the Hue, Saturation and Value channels;

3. The eigenvalue and eigenvectors of I_{ADF} are computed and a binary mask image B_M is composed;
4. An enhanced Hue channel image I_{EN} is obtained with contrast enhancement using B_M for the intensities adjustment;

- **Region growing process**

5. The binary mask B_M is used to place the seeds $\mathscr{S} = \{s_1, ..., s_N\}$ in I_{EN};
6. Pixel intensities from both alga and background regions in I_{EN} are sampled using the binary mask B_M;
7. The Gaussian probability distributions parameters $\theta_1 = \{\mu_1, \sigma_1\}$ and $\theta_2 = \{\mu_2, \sigma_2\}$ are estimated from the alga and the background region samples, respectively.
8. The Gaussian distributions P_1 and P_2 of those regions are computed according to Eq. (4.32).
9. Region growing is applied to image I_{EN} using the set of seeds $\mathscr{S}$, the probability distributions P_1 and P_2, resulting in a binary image B_{RG} composed by the set of alga regions and background;

- **Post-segmentation**

10. The binary image B_{RG} is used as input to the rolling ball morphological operation to obtain the image I_{SEG}, which is the final segmentation.

The image in Figure 45 shows an alga colony, which poses a particular challenging segmentation case. Because the organisms are typically moving when the digital image is captured, the color intensity of alga cells vary considerably. The original image is shown in Figure 45(a), in which the areas where the different cells meet are nearly transparent. Figure 45(b) shows the equalized image obtained during the contrast improvement of the Hue channel. The enhancement process produces the new Hue channel shown in Figure 45(c). The image resulting from the region growing algorithm is presented in Figure 45(d), and Figure 45(e) shows the final segmentation, in which several holes were successfully filled by the rolling ball operator. Figure 45(f) depicts the associated ground-truth. One notices that the segmentation has preserved very well the stellate shape aspect of the colony.

Another segmentation case concerns an elongated alga cell which shows mucilage in the bottom corner. Handling the mucilage properly is very difficult. Ideally, it should remain connected to the alga cells, as it is characteristic of this kind of algae, but conventional segmentation approaches will very likely separate the structures. Figure 46(a) depicts the original RGB image with a single alga and the seed point obtained (in red). Figure 46(b) shows the Hue channel, which is pretty noisy in this case. The smoothing and the enhancement processes produce an image with homogeneous regions and improved contrast between algae and background,

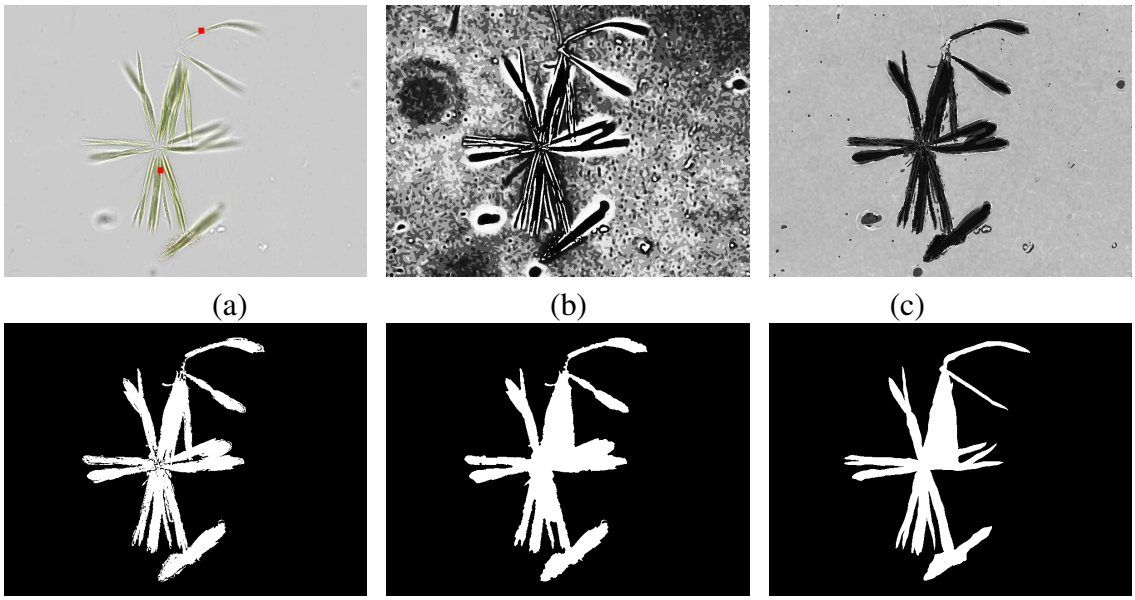


Figure 45 – Segmentation of an alga colony: (a) original RGB image (seeds denoted by the red points); (b) the equalized image I_{EQ} determined during the enhancement process; (c) enhanced Hue channel after contrast enhancement; (d) output of the region growing algorithm; (e) result after applying the rolling ball operator; (f) ground-truth image.

shown in Figure 46(c). The final segmentation is presented in Figure 46(d), and Figure 46(f) presents the corresponding segmentation obtained of the level set approach, which shows the alga organism without its distinguishing mucilage structure. The ground-truth is shown in Figure 46(f), and a comparison with Figure 46(d) again indicates a very good preservation of the alga shape including its particular mucilage.

Figure 47 illustrates an application of the proposed segmentation methodology to other types of digital microalga images that share similar characteristics to the Selenestraceae images. Figure 47(a) presents the original RGB image of *Micrasterias pinnatifida* alga, with seeds indicated by the red points. Figures 47(c) and 47(d) show the original Hue channel and the enhanced Hue images, respectively, obtained with the same enhancement process applied to the Selenastraceae images. Figure 47(d) depicts the final binary image obtained after the region growing and the rolling ball procedures. It can be noticed that the proposed region-growing methodology also works well on other types of alga images, but target cells should be green.

Section 3.5 presents the qualitative and quantitative results obtained from applying the proposed segmentation methodologies and others from the literature on a particular set of green alga images.

3.5 Experimental results

The performance and the effectiveness of the proposed segmentation methodologies have been evaluated on a set of 40 green alga images depicting different species of the Selenas-

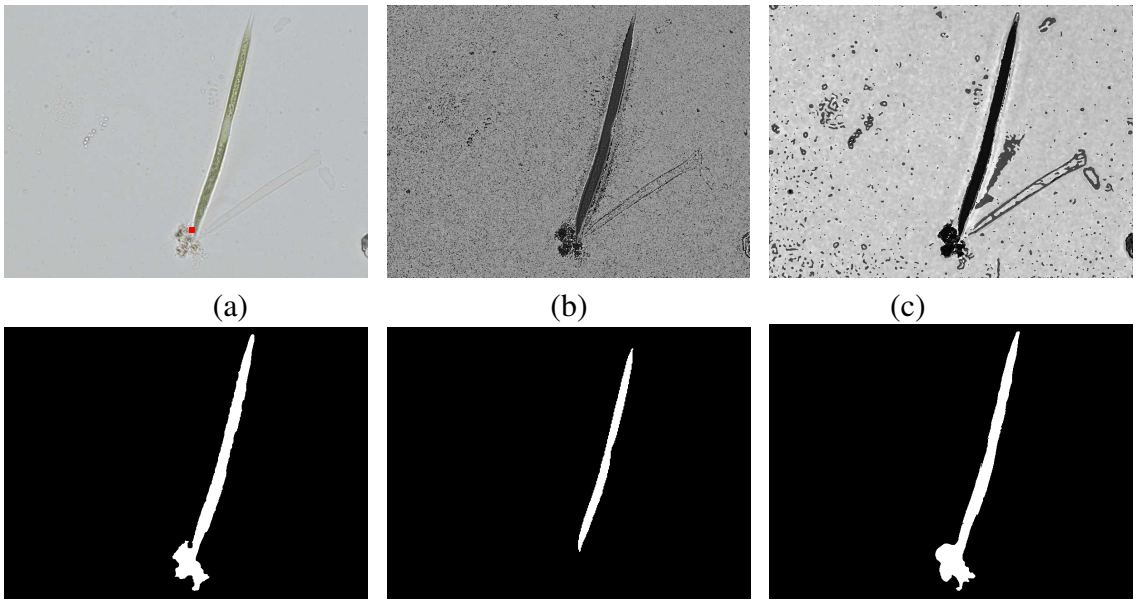


Figure 46 – Segmentation of an elongated alga cell: (a) original RGB image and its seed (in red) overplaced to the cell; (b) original Hue channel; (c) enhanced Hue channel; (d) segmentation after application of the rolling ball operator; (e) obtained result using the proposed methodology based on level set method; (f) ground-truth image.

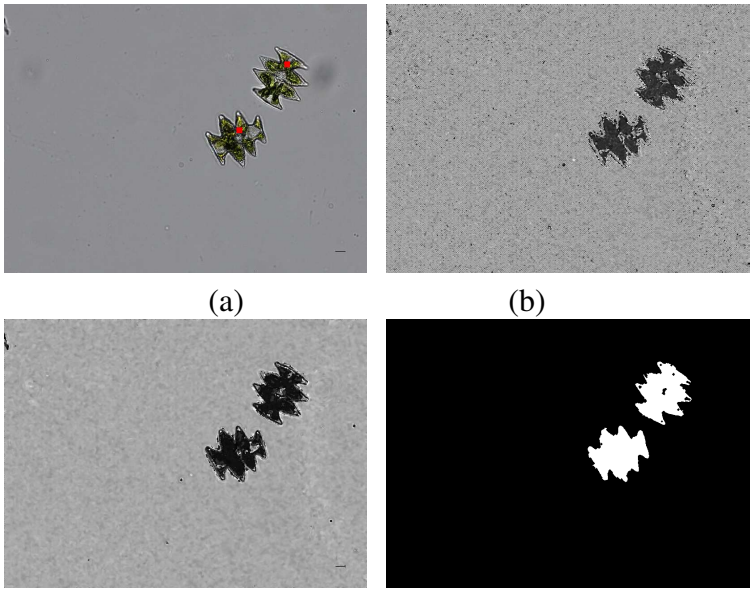


Figure 47 – Segmentation of *Micrasterias pinnatifida* in a microscopy image: (a) original RGB image; (b) original Hue channel; (c) enhanced Hue channel; (d) final segmentation.

traceae algae family complex. The experiments have been executed on a *Intel(R) Core(TM) i7 2.40GHz* and the proposed techniques have been implemented in Matlab.

Segmentation results obtained with the proposed methodologies and with methods from the literature are compared, in terms of accuracy, with manual segmentations of the images provided by an expert biologist, referred to as ground-truth (GT) images. Region accuracy is measured with the Jaccard coefficient [177] and the F-Measure [178]. The F-Measure (F_1) is

defined in terms of the precision (Pr) and the recall (Rc):

$$F_1 = 2\frac{Pr \;\; Rc}{Pr + Rc}. \tag{3.21}$$

Precision measures the percentage of region pixels in the automatic segmentation that correspond to region pixels in the ground-truth, being sensitive to over-segmentation. Recall measures the percentage of region pixels in the ground-truth that were detected via automatic segmentation, and is sensitive to under-segmentation. These measures are computed as:

$$Pr = \frac{TP}{TP + FP} \qquad Rc = \frac{TP}{TP + FN}, \tag{3.22}$$

where TP (true positive) refers to the pixels labeled as belonging to alga regions in both segmentation and GT. FP (false positive) refers to the pixels labeled as belonging to alga regions in the segmentation, but as non-alga pixels in GT. TN (true negative) refers to the pixels labeled as non-alga in both segmentation and in the GT. FN (false negative) refers to the pixels labeled as non-alga in the segmentation, but are actually pixels belonging to alga regions in the GT image. The F-Measure is the weighted average between precision and recall, in which F_1 values close to 1 indicate high segmentation accuracy and 0 indicates the lowest-possible segmentation accuracy.

The Jaccard coefficient (Jc) is defined as:

$$Jc = \frac{|I_S \cap I_{GT}|}{|I_S \cup I_{GT}|}, \tag{3.23}$$

in which $|\cdot|$ is the cardinality operator, I_S is the segmented image (binary image) and I_{GT} is the ground-truth image. $Jc = 1$ when there is an exact match between the segmentation and GT, and $Jc = 0$ when a complete mismatch is observed. For both Jaccard coefficient and F-Measure, the average accuracy for an image set is computed by averaging the accuracy values computed for each image.

The proposed segmentation methodologies are initially compared with three other possible approaches: two of them are techniques commonly employed for segmenting biological images, namely the thresholding-based binarization with Otsu's automatic method for computing the threshold value, and the Watershed transform.

To ensure a meaningful comparison, the images input to the four segmentation approaches are the smoothed original images resulting from applying the ADF filter with the same parameter settings, and the rolling ball transformation is applied to the initial segmentation results in all cases. The segmentations obtained with the Watershed transform required some additional postprocessing, as the method outputs multiple subareas. The relevant alga region is selected from the set of subareas by determining the seed points with the approach described in Section 3.4.2 and identifying the subareas that include the seed points.

Table 1 presents the computed Jaccard (Jc) coefficients and F-measures (F_1) for the proposed segmentation methodologies and the comparison techniques for the test image set. Moreover, the average running times (in seconds) are presented to compare the practical performance of the region growing based approach against the segmentation solution based on the level set approach. The region growing methodology yielded higher accuracy rates. This is mainly due to the appropriate design of the preprocessing and postsegmentation steps, as well as the usage of Gaussian distributions to characterize the target regions.

Table 1 – Average accuracy rates and execution times.

Segmentation techniques	Jc	F_1	time (s)
1. Methodology based on region growing	**0.77**	**0.88**	49.0
2. Methodology based on level set	0.72	0.82	253.2
3. Binarization using Otsu's threshold	0.44	0.55	**1.2**
4. Watershed transform	0.53	0.61	22.1

The running times of each technique indicate that the level set based approach is computationally more expensive, which can be explained by the underlying optimization process required to compute the numerical solution of the PDEs associated to the level set equation. In general, if a level set curve converges after k iterations, the overall algorithm complexity is denoted by $O(m_1 m_2 k)$, in which m_1 and m_2 are the dimensions of the input image. On the other hand, as the region growing methodology follows the implementation proposed by Gonzalez et al. [174], the complexity of the growing process is $O(N m_1 m_2)$, in which N is the number of seeds. As k is likely to be higher than N, the complexity of the level set approach is higher when compared to the adopted implementation of the region growing algorithm.

A particular segmentation case is shown to illustrate the result obtained with the proposed methodologies and compare them with those obtained with the Otsu's threshold binarization and the Watershed transform. The original image, depicting a single alga and its computed seed point (in red), is illustrated in Figure 48(a), whereas Figure 48(b) shows the manual segmentation generated by the biologist. Figure 48(c) shows the segmentation result obtained with the region growing based method. Figures 48(d), 48(e) and 48(f) illustrate the results obtained, respectively, with the binarization approach, the adapted Watershed transform and the proposed level set methodology.

The segmentation obtained with the proposed region growing strategy is clearly most similar to the ground-truth, with segmentation accuracies given by $F_1 = 0.90$ and $Jc = 0.82$. The segmentation obtained with the thresholding-based method deformed the original alga shape (accuracy rates are $F_1 = 0.71$ and $Jc = 0.55$). The Watershed-based method achieved accuracy rates $F_1 = 0.70$ and $Jc = 0.75$ in this case, but the binary region presents a rougher contour, even though the shape properties are well preserved. The proposed level set methodology also output a highly accurate segmentation ($F_1 = 0.88$ and $Jc = 0.79$), but still inferior to the region growing, besides being more time consuming.

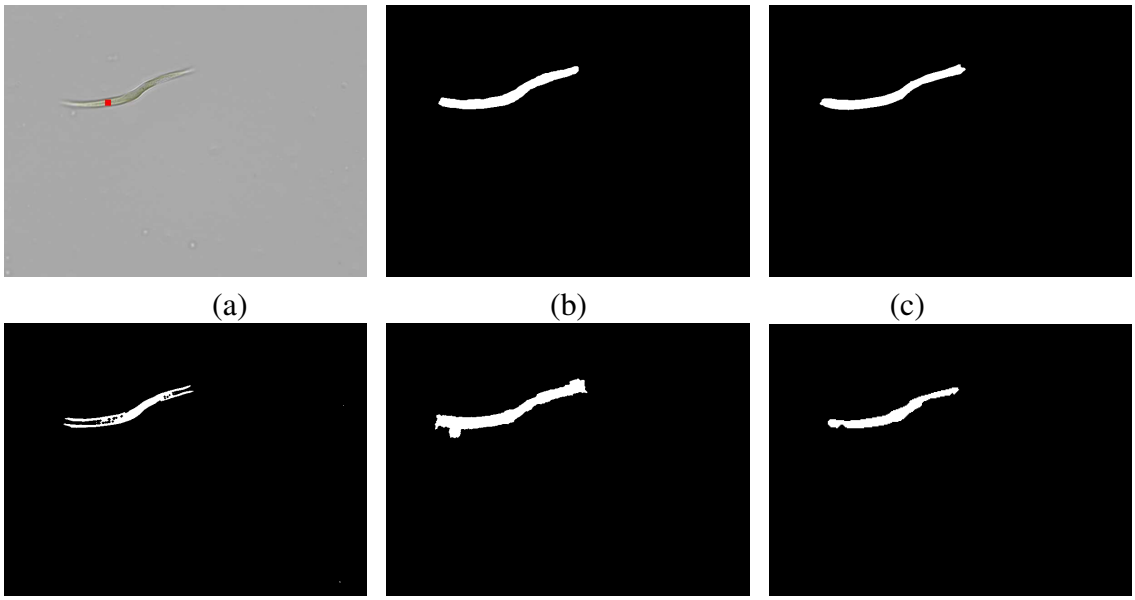


Figure 48 – Comparing results from different segmentation techniques: (a) original RGB image with seed (in red) overlapped to the cell; (b) ground-truth image; (c) segmentation obtained with the proposed region growing method; (d) segmentation with thresholding; (e) segmentation with the Watershed method; (f) segmentation with the proposed level set method.

Given the running time performance and the segmentation accuracy, the method based on region growing is chosen as the best alternative for segmenting alga shapes, since a highly accurate segmentation is extremely relevant for the subsequent feature extraction step. Moreover, efficiency in the segmentation process is also important in a real-time application. The good performance is a result of the high specialization of the region growing methodology, including the steps of image enhancement and post-processing that preserve the overall shape geometries. However, the strategy for determining seeds could not detect all alga regions, although it ensures that at least one alga region is found from the mask.

On the other hand, the segmentation methodology based on level set showed to be more robust since it can identify all alga regions and handle the smooth intensity variation in background that can influence the final segmentation. However, the segmentations obtained are not as precise for individual alga regions, because small cell concavities and noisy boundaries may be identified due to transparencies.

3.6 Final considerations

Segmentation of green alga shapes from the available images is particularly challenging due to complex image characteristics such as low contrast, non-uniform illumination conditions across images and blurred alga boundaries resulting from alga movement during image acquisition. Traditional segmentation techniques fail when attempting to identify and detect such microorganisms in digital images. Therefore, this chapter introduced two segmentation

methodologies for accurate segmentation of green microalga shapes, so that precise shapes can obtained for the subsequent feature extraction.

The first technique relies on a level set approach combined with a Bayesian formulation. The image regions, i.e., the alga cells and the background, are described by multivariate Gaussian distributions computed prior to the curve evolution process, estimated by means of region intensity samples. The level set evolves from an initial position towards the alga cells, considering *a priori* region properties, such as intensity variation and texture, and edge information for shape preservation. In this method, the level set formulation incorporates an edge potential function to preserve boundaries during the segmentation.

The second methodology employs a region growing principle that incorporates specific smoothing and contrast enhancement steps. In order to handle transparencies and blurred intensities in alga cells, the HSV model is employed for alternatively representing the intensities of image regions. As in the level set method, the alga and background regions are described by Gaussian distributions using intensity samples from the Hue channel. Seed points associated with specific regions are computed and the growth process groups neighboring pixels that satisfy a predefined homogeneity criterion derived from the Gaussian distributions.

Experimental results have shown that the proposed methodologies achieve high segmentation accuracies when compared with ground-truth segmentations provided by the biologists. Moreover, it also yielded better accuracy rates than existing methods from the literature, such as segmentation with the Watershed transform and the binarization using Otsu's technique.

As described in the following Chapter, resulting shapes have been input to shape feature extractors that can produce effective shape descriptors for distinguishing between multiple alga species.

CHAPTER

4

SHAPE-BASED FEATURE EXTRACTION

The previous chapter presented two methodologies for segmenting green alga images accurately. Those methods produce binary images of the alga silhouettes identified. However, providing the segmented images directly as input to the classifiers or visualization techniques would be ineffective due to a large amount of redundant information. A suitable strategy is to compute a compact set of statistical measures describing the relevant image features.

Typically, manual identification of species is performed observing the alga morphological characteristics in images, which is time consuming and unfeasible as image sets get larger. Such practical limitations motivate the development of automatic methodologies for computing measures associated with alga morphological characteristics.

This chapter describes studies on shape feature extraction from 2D binary images, focusing on the representation and description of green algae. Section 4.1 describes the shape representation adopted to compute the relevant features of green microalgae. Section 4.2 presents some basic shape geometric features extensively used in the literature and other simple methods for shape description. Section 4.3 proposes a general shape descriptor named Segment Intersection Descriptor (SID) that can be applied to shapes of different natures. Section 4.4 describes two green alga descriptors proposed in this research, in which one is based on basic shape features and the other relies on the SID. Section 4.5 reports the experiments performed on green alga shapes and benchmark shape sets to evaluate the proposed shape descriptor and the specific green alga descriptors. Section 4.6 discusses the contributions and preliminary results concerning the proposed shape descriptors.

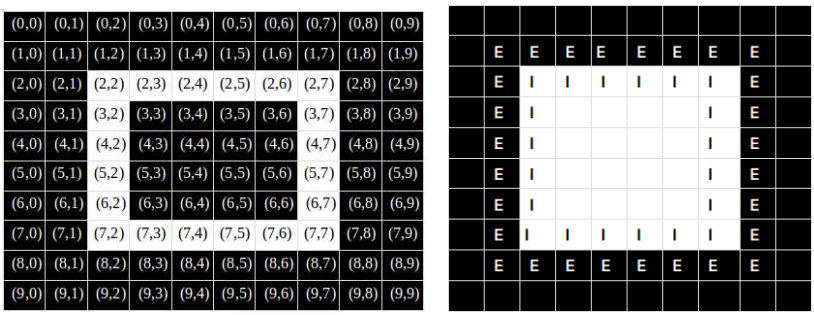
4.1 Shape representation

A shape in a binary image can be represented in terms of its boundary, or in terms of the pixels within the associated region [64]. External representations are more appropriate to characterize boundaries and internal representations require a set of connected pixels defining

the shape. Nonetheless, feature extractors can also combine both representations for a complete shape description.

Shape contour

An intuitive strategy for external representation is the parametric contour representation, given by the absolute coordinates of the boundary pixels. A parametric representation of a region boundary [86] is defined as $\mathscr{C}(t) = (x(t), y(t))$, for $t \in [0,1]$, discretized as $C = \{(x_1, y_1), ..., (x_i, y_i), ..., (x_{n_p}, y_{n_p})\}$ for $1 \leq i \leq n_p$. Figure 49(a) illustrates the parametric representation, in which the shape contour is given by the white cells and can be defined as $\{(2,3),(2,4),(2,5),(2,6),(2,7),(3,7),(4,7),(5,7),(6,7),(7,7),(7,6),...,(6,3),(5,3),(4,3),(3,3)\}$. An important aspect of contour representations is to consider whether to adopt an external or an internal shape contour. Figure 49(b) illustrates both internal (I) and external (E) contours, in which the first internal approach is formed by the contour points belonging to the object, while the second is defined by the neighboring points surrounding it. In this work, the clockwise direction and the internal approach are chosen for analysing and representing shape contours.



(a) (b)

Figure 49 – Contour representation: (a) Shape contour defined by the white cells in a parametric form; (b) Contiguous points of internal (I) and external (E) approaches for contour representation.

Boundary-following algorithms [179] [180] track region boundaries generating as output an ordered sequence of boundary points. Such algorithms require defining the direction of the pixel-by-pixel tracking process (clockwise or counterclockwise) and a "first" contour point, which indicates where the tracking procedure starts and finishes. Assuming that white (one-valued) pixels describe the objects (or regions) and black (zero-valued) pixels are the background, the tracking procedure visits consecutive pixels of the shape contour until returning the the initial pixel, indicating the end of this procedure.

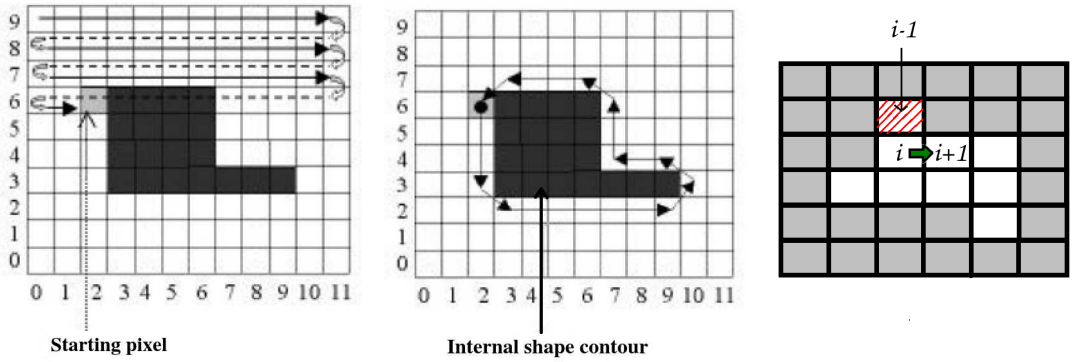
In this work, the Moore boundary-tracking algorithm [64] is applied to obtain a parametric representation of the internal shape contour. The first pixel is determined by analysing the image pixels from left to right and from top to bottom until reaching the first white pixel, as shown in Figure 50(a). In order to avoid erroneous tracking caused by incomplete spurs removal, the background point b_0 to the left of the starting point p_0 is stored, to prevent the algorithm returning to the starting point without analysing all the contour points. Then, the

boundary-following algorithm occurs in the direction shown in Figure 50(b) and is described by the following steps:

1. For p_0 and b_0, examine the 8-neighborhood of p_0, starting at b_0 and proceeding in a clockwise direction. Let p_1 be first one-valued neighbor pixel encountered and assign b_1 as the background point that precedes p_1. Store the locations of a_0 and a_1 to be used as a stopping criterion.

2. Let $p_i = p_1$ and $b_i = b_1$. Analyse the 8-neighborhood of p (denoted as $n_1, ..., n_8$) and identify the first n_k labeled as one (belonging to the contour). Do $p = n_k$ and $b = n_{k-1}$;

3. Repeat Step 2 until $p = p_0$ and the next boundary point to be visited is p_1.



(a) (b) (c)

Figure 50 – Algorithm for extracting the internal shape contours in binary images: (a) Searching the starting point (first pixel); (b) The tracking procedure through the contour points in the adopted direction; (c) Tracking procedure and decision about the next pixel to visit (green arrow).

After finding the first pixel p_0, the algorithm chooses the next pixel to visit according to the order in which the neighbors are analysed. Following clockwise from reference point b_0, the next pixel to visit is the first found in the object boundary. The current point is now p_1 and b_1 is p_0. Figure 50(c) illustrates this operation, where p_i is the current pixel, p_{i-1} is the previous visited pixel (with red dashes), p_{i+1} is the next pixel to visit (indicated by the green arrow). The direction between the visited pixel and the current pixel is updated to determine the next point, preventing an infinite loop or a return to already visited points.

Signatures

A shape signature is a 1-D function representing 2-D areas or boundaries, usually describing a unique shape and capturing the perceptual property of the shape. The basic idea is to reduce the boundary representation to a 1-D function, which might be easier to describe in a compact manner than the original 2-D boundary. Generally, 1-D functions are used as a preprocessing step for feature extraction since new statistical measures can be derived from that representation.

The curvature function is a well-known representative of 1-D functions. Curvature is a very intuitive property to the human perception, as it is a relevant boundary feature for humans to judge similarity between shapes [112] [181]. Shape segments with a straight pattern have curvature values close to zero, while segments close to corner areas or with irregularities have higher curvature values. Eq. (4.1) shows how to compute the curvature value in a given point (x_i, y_i):

$$\kappa_i = \frac{x'_i y''_i - x''_i y'_i}{((x'_i)^2 + (y'_i)^2)^{3/2}}, \tag{4.1}$$

in which $x'_i = x_{i+1} - x_i$ and $y'_i = y_{i+1} - y_i$ are approximations to first order derivatives to x and y, respectively. Similarly, $x''_i = x_{i+1} - 2x_i + x_{i-1}$ and $y''_i = y_{i+1} - 2y_i + y_{i-1}$ are second order derivative approximations to x and y. For the curvature measure to be invariant to scale, Eq. (4.1) is normalized by the absolute mean curvature:

$$\kappa'_i = \frac{\kappa_i}{\frac{1}{n_p}\sum_{i=1}^{n_p} |\kappa_i|}, \tag{4.2}$$

in which κ'_i is invariant to rotation, translation and scale. Figure 51 illustrates two curvature functions obtained from distinct shapes. Different patterns are noticeable in their signature functions, from which relevant statistical measures can be thus computed. An example is the average bending energy, which is given by Eq. (4.3):

$$ABE = \frac{1}{n_p}\sum_{i=1}^{n_p} |\kappa'_i|^2. \tag{4.3}$$

However, it is worth noting that curvatures can be degraded by noise or aliasing effects when the shape contour is subsampled.

Skeleton

The skeleton is an effective representation that reveals the essential structure of a shape. Several approaches have been proposed for extracting skeletons from binary regions in images. A popular solution is the medial axis transform (MAT) [182] which combines, in a unique way, local boundary information with local region information [183]. The key idea of MAT relies on computing the nearest distance from each interior point to the boundary, such as the distance transform of a shape. Internal points belong to the skeleton if they have at least two nearest points to the boundary. Then, skeletons are defined as the higher responses on the distance field. Despite its simple formulation, MAT tends to generate medial axis with unwanted spurs, which may interfere with recognition processes [184].

In this work, the multiscale approach based on the Image Forest Transform (IFT) proposed by Torres and Falcão [114] is employed to compute skeletons. In the IFT algorithm, each

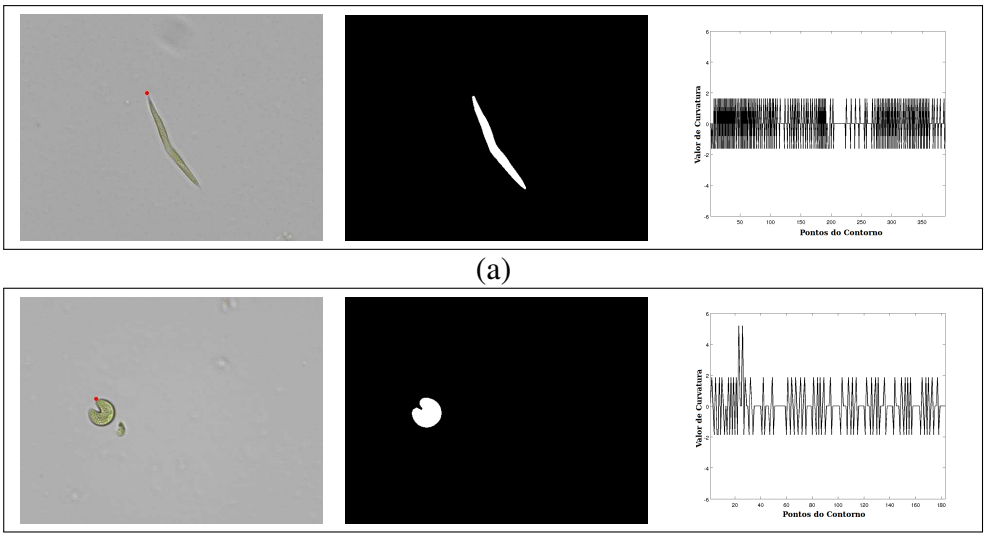


Figure 51 – Curvature signature: comparing two alga signatures (a) and (b): original images and the starting point (x_0, y_0) denoted by the red point, corresponding alga silhouette, and the normalized curvature function.

contour point is taken as a seed and from these points a minimal cost path is determined. Each path is a tree and their union is a forest that covers the entire image. The IFT produces a label map, which allows partitioning the image into discrete Voronoi regions. A difference image is obtained by extracting the boundaries of these Voronoi regions and then thresholded to obtain the internal skeleton.

Figure 52(a) illustrates an arbitrary green alga shape. Figure 52(b) depicts the Voronoi diagram representing the label map given by IFT. Figure 52(c) presents the computed skeleton (the white axes) overlapping the alga shape after thresholding the difference image of the Voronoi diagram regions. Some small spurs can be noted at the alga extremities in Figure 52(c).

Figure 52 – Skeletons: (a) Green alga shape; (b) the Voronoi diagram; (c) the alga shape and the skeleton (in white) subscripted to it.

The skeleton representation is invariant to rotation, translation and scale if the shape contour is normalized to $[0, 1]$. Skeletons are known to be highly sensitive to deformations in object boundaries, which requires pruning the undesirable branches [185]. This task is quite challenging due to the complexity in defining appropriate criteria to maintain only the relevant branches in the final skeleton.

4.2 Basic shape description

After adopting a suitable representation for shapes, subsequent steps concern computing shape measures to be used as features for classification and pattern recognition. Several techniques have been employed to extract relevant information on the visual properties of a shape.

The basic geometric features are simple measures related to metric aspects of a shape [186]. Generally, these measures are capable of distinguishing between clearly different shapes, but they do not describe uniquely the visual properties of a shape [115]. However, they are combined with other shape descriptors to improve the discriminability among shape classes. Some basic geometric features are described below:

Perimeter (*Pe*): is the arc length of a spatial curve (or contour). Considering a discrete representation of a binary contour, this parameter is given by the number of points (pixels) in the contour. In this work, a 8-neighborhood connectivity is adopted when computing the perimeter value. In a parametric representation, the perimeter of a shape S is given by Eq. (4.4):

$$Pe = \int \sqrt{x^2(t) + y^2(t)}dt \qquad t \in [0,1]. \tag{4.4}$$

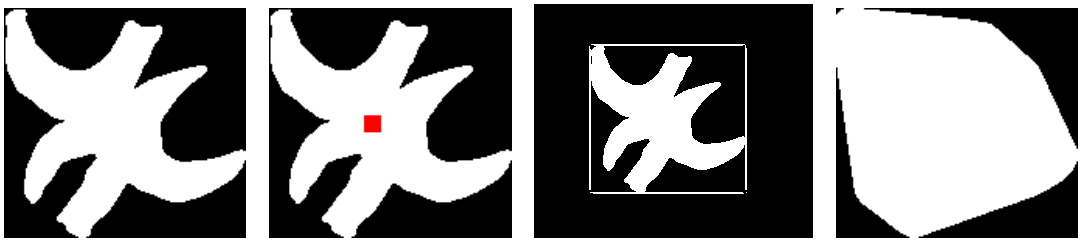
This measure is invariant to rotation and translation, but not to scale.

Area: is a measure of the total spatial occupation of a shape, given in number of pixels. It can be estimated by counting the number of one-valued points in the binary image. Eq. (4.5) estimates the shape area assuming a shape isolated into a binary image (or in its bounding box):

$$Area = \frac{1}{2}|\sum_{i=0}^{n_p-1}(x_i y_{i+1} - x_{i+1} y_i)|. \tag{4.5}$$

Area is invariant to rotation and translation, but not to scale.

Diameter: also known as the maximum chord, it is the largest pairwise distance in a shape. Although a brute force algorithm could be employed, this value can be estimated from the signal described in the complex domain to reduce the computational time. Assume the complex signal $u = b + ci$, in which b an c are the spatial coordinates of the boundary points. Algorithm 1 describes the steps to compute the diameter of a shape given by a



(a) (b) (c) (d)

Figure 53 – Basic geometric features: (a) Alga shape in a bounding box; (b) Shape centroid (in red); (c) Minimum bounding rectangle; (d) Convex hull.

contour with n_p points [186]:

Algorithm 1: Algorithm for computing the shape diameter.

Data: Shape contour points $\{p_1, ..., p_{n_p}\}$

Result: Shape diameter

```
1 d_max ← 0;
2 for j1 ← 1 TO n_p − 1 do
3     for j2 ← j1 + 1 TO n_p do
4         if |u(j1) − u(j2)| > d_max then
5             d_max ← |u(j1) − u(j2)|;
6             lm = j1;
7             cm = j2;
```

The diameter is invariant to rotation and translation, but not invariant to scale.

Centroid: the shape centroid (c_x, c_y) is calculated using the discrete contour points as in Eq. (4.6):

$$c_x = \frac{1}{n_p}\sum_{i=1}^{n_p} x_i \qquad c_y = \frac{1}{n_p}\sum_{i=1}^{n_p} y_i. \tag{4.6}$$

It is worth noting that the shape centroid may be placed outside the shape. To overcome this situation, a skeleton-based approach is employed to obtain a centroid that is inside the shape. The shape centroid is assigned as the median value among the spatial coordinates of all skeleton points. The shape centroid is invariant to scale, rotation and translation. Figure 53(b) illustrates the shape centroid, denoted by the red point.

Minimal Bounding Rectangle (MBR)**:** also known as bounding box or enclosing rectangle [187], the minimal bounding rectangle is the smallest rectangle with an arbitrary orientation, that contains the shape, as depicted in Figure 53(c). In this work, the technique proposed by Chaudhuri et al. [188] is employed to compute the minimum enclosing rectangle, described by Algorithm 2. The algorithm assumes that the major axis and the

minor axis of the shape cross the centroid and are perpendicular to each other. Such axes are compute using a least-square approach also detailed in their work.

Algorithm 2: Algorithm for computing the minimum bounding rectangle according to the approach described by Chaudhuri et al. [188].

Data: Shape contour points $\{p_1, ..., p_n\}$

Result: The vertices of the bounding rectangle

1 Compute the shape centroids;
2 Determine the principal shape axis that provides the shape orientation;
3 Compute the upper and lower edge points to the shape's major and minor axis;
4 Obtain the four vertices of the bounding rectangle using parallel lines derived from the upper and lower edge points;

Eccentricity: provides the ratio of the length of the major axis to the length of the minor axis. Such axes can be obtained with to the eigen-axes method or between the minor and major axes of the ellipse surrounding the shape. The MBR can be used to compute the eccentricity by determining the rate between the higher bound L_{MBR} and the lower bound W_{MBR}:

$$Ecc(W_{MBR}, L_{MBR}) = \frac{W_{MBR}}{L_{MBR}}. \tag{4.7}$$

Convex Hull ($\mathscr{H}$): the convex hull $\mathscr{H}$ of a shape S is the smallest convex set containing S, which can also be denoted as a polygon. The convex hull is useful for object recognition and to compute other shape parameters. Several algorithms have been proposed to compute the convex hull, such as morphological operation [64] or monotone hulls [189]. The monotone convex hull sorts the points according to a predefined ordinate (x or y) and then constructs upper and lower hulls of the points. Figure 53(d) presents the convex hull of the alga shape in Figure 53(a).

Rectangularity: indicates how much the shape fills its minimum bounding rectangle:

$$Rect = \frac{A(S)}{A(MBR(S))}, \tag{4.8}$$

in which $A(MBR(S))$ is the area of the minimum bounding rectangle.

Convexity: is defined as the ratio of the convex hull perimeter in relation to the shape perimeter:

$$Cx(S) = \frac{Pe(\mathscr{H}(S))}{Pe(S)}. \tag{4.9}$$

Solidity: describes the extent to which a shape is convex or concave [190]. It measures the ratio of the area of the convex hull $A(\mathscr{H}(S))$ and the shape area $A(S)$.

$$Sol(S) = \frac{A(S)}{A(\mathscr{H}(S))}. \tag{4.10}$$

Thickness: provides a quantitative value for the thickness of a shape. This parameter is computed by means of a morphological operation, which performs successive erosions in a shape and counts the number of required operations until it shrinks completely. Algorithm 3 describes the process, in which higher values of th indicate thick shapes, while lower values denote thin shapes.

Algorithm 3: Compute a measure of the thickness of a shape.

Data: Binary shape S
Result: Thickness value th

```
th ← 0;
S' ← S;
str ← structuring element (disk, radius length = 1);
while Area(S') > 0 do
    S' ← erosion on S' using str;
    th ← th + 1;
```

It is common to use only a subset of the basic geometric features to describe the morphological characteristics of microalgae in digital images [191] [9]. Some authors emphasize that those descriptors are useful to categorize simple shape sets, or be combined with other descriptors if required [115] [181]. The Selenastraceae alga shapes are complex because distinct species are visually similar, demanding a more sophisticated formulation that captures relevant shape patterns for an appropriate description.

Therefore, this work adopts the strategy of combining basic geometric features with features obtained from more elaborate descriptors in order to achieve a powerful description and high discriminability among alga species. The next section presents a descriptor useful to characterize different kinds of shape, which is invariant to translation, and presents low variance to scale and rotation.

4.3 Segment Intersection Descriptor

The Segment Intersections Descriptor (SID) represents the patterns and geometry of a shape as a signature function. This descriptor combines region and contour information into compact signature representations. Several qualitative, quantitative and some statistical parameters are obtained from the signatures to measure the shape's internal and external properties.

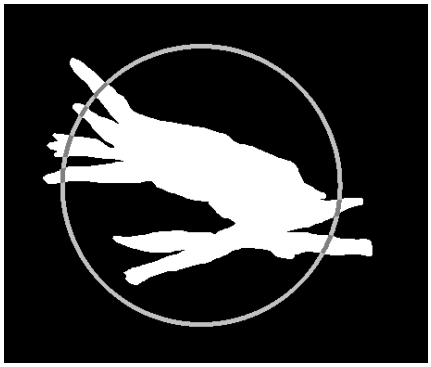


Figure 54 – Step of the circumference growth process: the overlapping pixels between the circumference and the shape form segments which are denoted by the sets of connected dark gray pixels.

The rationale of SID consists in growing a circumference from the shape's centroid by varying the radius length and seeking for its overlapping pixels to the shape. The process stops when no overlapping pixels are identified between the circumference and the shape. Figure 54 illustrates the circumference (in light gray intensities) at a specific step, in which the overlapping pixels with the alga shape are denoted by the dark gray intensities. Such pixels define circumference segments that intersect the shape, forming the segments, a set of connected pixels with gray intensities. At each step of the circumference growth, two measures are computed: the number of segments (Figure 54 presents 6 segments) and the rate of overlapping pixels with respect to the circumference, computed according to Eq. (4.11):

$$\text{Rate of overlapping pixels} = \frac{\text{\# dark gray pixels}}{\text{\# dark gray pixels} + \text{\# light gray pixels}}. \tag{4.11}$$

As these two measures are computed at each step of the circumference growth, it is straightforward to derive signatures.

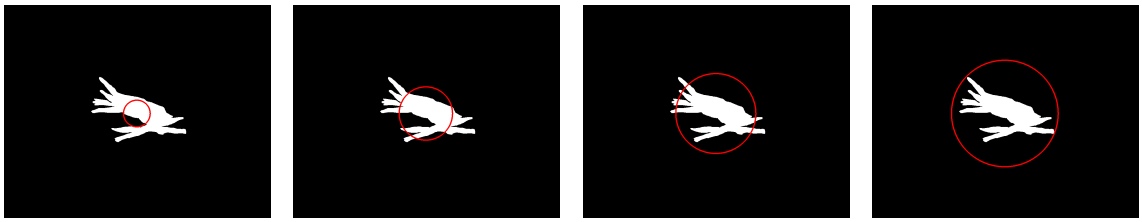
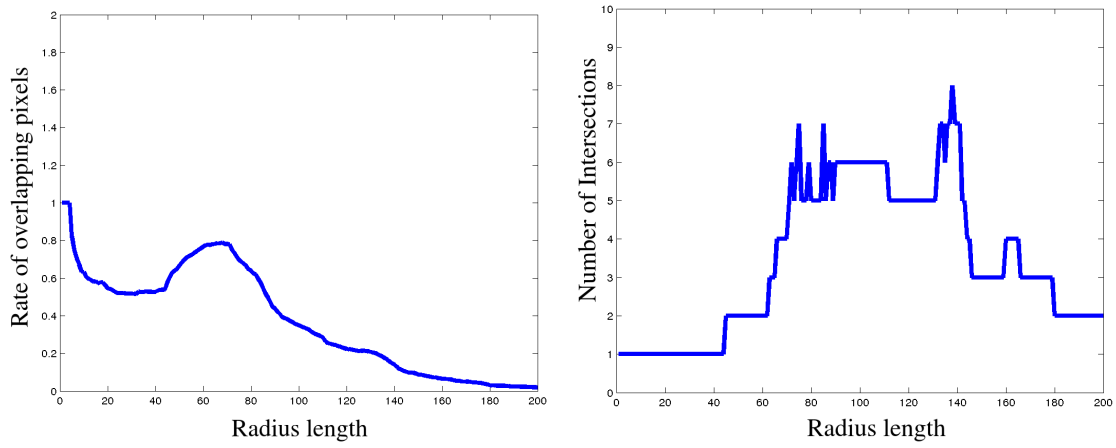


Figure 55 – Obtaining the signature representation: circumference growing for radius sizes 50, 100, 150 and 200, respectively.

The procedure to computing the signature functions is simple and intuitive. Given the shape's centroid, the process starts with a unitary radius and computes the rate of circumference points that intersect the shape. Obviously, in the first iterations, this rate is usually one, since the whole circumference is inside the shape. The process is then repeated with increasing radius values, plotting the circumference and computing the rate of intersections (Eq. 4.11) and the number of segments. Circumference growing stops when no more intersections are detected,

i.e., the circumference is totally outside the shape. Figure 55 depicts this process for radius sizes 50, 100, 150 and 200.



(a) (b)

Figure 56 – Signature functions obtained from the process in Figure 55: (a) rate of intersecting points between the circumference and the shape; (b) number of intersections between the circumference and the shape.

As a result, two signature functions are obtained, shown in Figure 56: one is referred as the *Rate of overlapping pixels* $\times$ *radius length*, shown in (a), and the other as *Number of intersecting segments* $\times$ *radius length*, shown in (b). Such signatures were derived from the circumference growth process depicted in Figure 55. It is expected that shapes in the same category produce similar signatures, while distinct shapes produce distinctive signatures.

Algorithm 4 describes the step-by-step process of growing the circumference and computing the number of segment intersections and rate of overlapping pixels signatures:

Algorithm 4: Segment Intersection Descriptor

Data: Binary shape S
Result: Two signatures nsi and ovp

$r \leftarrow 1$;
$(c_x, c_y) \leftarrow$ shape's centroid;
while $n_r > 0$ **do**
    Plot a circumference with radius r centered in (c_x, c_y);
    $nsi(r) \leftarrow$ number of segments with length greater than 2 that intersect the circumference;
    $n_r \leftarrow$ number of circumference pixels lying in the shape;
    $ovp(r) \leftarrow \frac{n_i}{Pe(Cir)}$;
    $r \leftarrow r + 1$;
$r_{max} \leftarrow r$;

For the sake of comparison, another example of the circumference growing is shown in Figure 57. The sequence of images in Figures 57(a), 57(b), 57(c) and 57(d) depict the plotted circumferences for radius values 1, 21, 41 and 62, respectively. Figures 57(e) and 57(f) plot the signatures, which are very different than those of Figure 56, in the length, the maximum

number of intersections and the rate of pixels intersecting the shape along the various radius length variation.

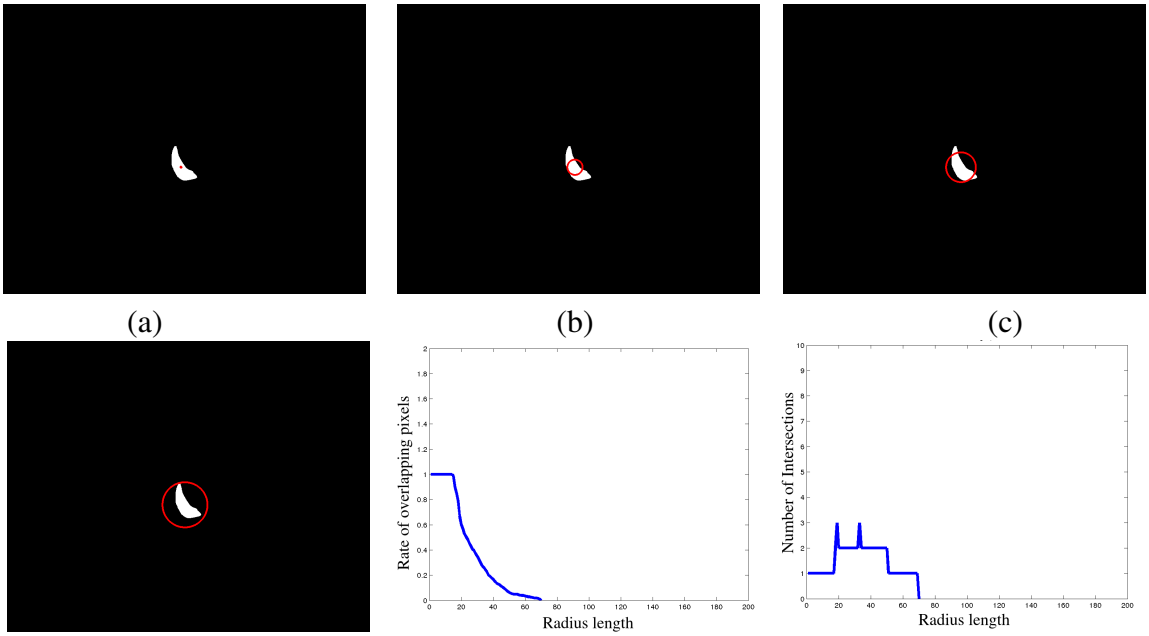


Figure 57 – Circumference growth in the shape of a sample of *Selenastrum bibraianum*: (a-d) growing the circumference by varying the radius values as, 1, 21, 41 and 62, respectively; (e) the rate of overlapping pixels signature; (f) the number of segment intersection signature.

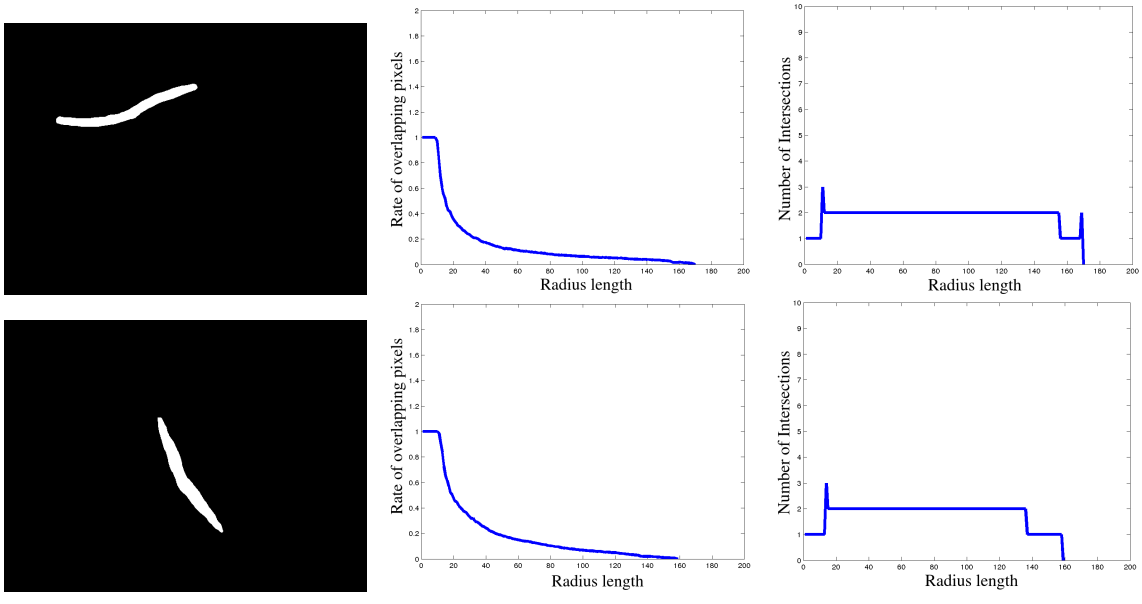


Figure 58 – Comparing the shape signatures of samples from the same species: (a) original binary shapes; (b) the rate of overlapping pixels signature; (c) the signature describing the number of segment intersection along the radius size variation.

Figure 58 compares different signatures obtained from alga shapes of the same species. The images in Figure 58(a) illustrate alga shapes of *Ankistrodesmus densus*, which are elongated and sometimes curved. A visual inspection allows to infer that the overall patterns of both

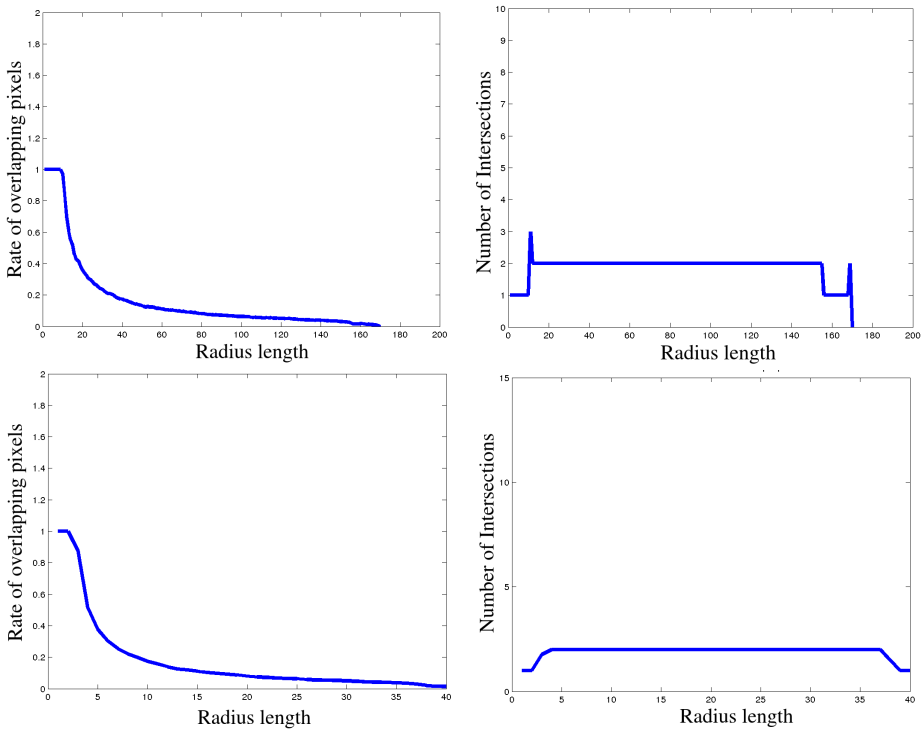
signatures are similar. The images in Figure 58(b) show the signatures of the rate of overlapping pixels and Figure 58(c) presents the signatures of the number of intersections.

During the circumference growth, two markers denoting radius length important to characterize the shape are selected from which additional interesting features are computed. The first marker r_{m_1} refers to the first radius length where the circumference is no more inside the shape. The other marker r_{mid} is computed as:

$$r_{mid} = \frac{r_{max} + r_{m_1}}{2} \tag{4.12}$$

in which r_{max} is the maximum radius size of the signature. In this approach, the shape centroid is obtained from the skeleton representation since it is guaranteed that the centroid is inside the shape. Moreover, segments with length less or equal than 2 are disregarded since they are formed by few pixels, which is not sufficiently meaningful to be considered a segment.

As shapes from different categories vary in size, their respective signatures have different lengths. Therefore, in order to reduce the descriptor's variance to scale and make it more robust to noise, both signatures are subsampled to a selected number n_l of equally spaced radius values, starting from radius size 5. The goal is to compose new signatures with reduced length, but preserving the original patterns.



(a) (b)

Figure 59 – Comparing the subsampling procedure considering the alga shape at the top in Figure 58(a): (a) original signatures of the rate of overlapping pixels (top) and number of intersections signatures (bottom); (b) respective subsampled signatures of the rate of overlapping pixels (top) and number of intersections signatures (bottom).

The choice of n_l depends on the resolution of the binary image, because shapes from

the same category but different scales should present similar signature patterns. Moreover, the choice of n_l must take into account the image resolutions of all shapes in a set. Using the available green alga images, several subsampled signatures were obtained by varying n_l from 10 to 60. Then a visual inspection by checking if the original and subsampled signatures presented similar patterns, n_l was set as 40. Figure 59 illustrates the subsampling procedure using the alga shape at the top in Figure 58(a) as reference. Figure 59(a) shows the original signature of the rate overlapping pixels in the top and its correspondent subsampled signature in the bottom. Figure 59(b) presents the original signature of the number of intersections and its subsampled representation. It can be seen that the signature patterns were preserved when compared to their original representations.

An advantage of subsampling all shape signatures to equal lengths is the possibility of using simple metrics to measure their pairwise dissimilarity. However, setting an inappropriate value for n_l leads to non-representative subsampled signatures that impair further recognition and classification steps. Using the entire signatures as feature vectors can lead to poor performance in shape classification since small shape variations produce subtle responses in their respective signatures (see the signatures of shape samples from the same class in Figure 58). Such variations influence the comparison of signatures when computing the dissimilarity, so that shapes in the same category are sometimes perceived as dissimilar. Thus, computing qualitative, quantitative and statistical measures from the signatures is a more robust solution.

Let $ovp(r)$ and $nsi(r)$ be two subsampled signature functions, designed to be the rate of overlapping pixels (*ovp*) to the circumference and the number of segment intersections (*nsi*), respectively, and r is the radius parameter. The quantitative metrics computed from these signatures are:

1. Maximum number of intersections

The maximum number of intersections among all radius values can differentiate between simple and complex geometric shapes, such as stellate or round algae. This measure is computed by Eq. (4.13):

$$maxNsi = \max_{r=1}^{r_{max}} \Big(nsi(r) \Big). \tag{4.13}$$

2. Average number of intersections

The average number of intersections among all radius values is less robust to noise than measure (1.) and is described by Eq. (4.14):

$$avgNsi = \frac{\sum_{r=1}^{r_{max}} nsi(r)}{r_{max}}. \tag{4.14}$$

3. Variance of the number of intersections

This measure describes the variation among the number of intersections for all radius values and is given by Eq. (4.15):

$$varNsi = \sum_{r=1}^{r_{max}} \left(nsi(r) - avgNsi \right)^2. \tag{4.15}$$

4. Average rate of overlapping pixels

The average signature rate of overlapping pixels between the circumference and the shape measures the homogeneity of a shape. For example, if an alga shape has a concavity, this measure can capture and express this pattern along the variation of the circumference radius. This measure is given by Eq. (4.16):

$$avgOvp = \frac{\sum_{r=1}^{r_{max}} ovp(r)}{r_{max}}. \tag{4.16}$$

5. Variance of the rate of overlapping pixels

This measure describes the variation among the rate of overlapping pixels for all the radius sizes and is given by Eq. (4.17):

$$varOvp = \sum_{r=1}^{r_{max}} \left(ovp(r) - avgOvp \right)^2. \tag{4.17}$$

6. Average length of segments

This parameter $(avgSizeNsi)$ computes a property of the length of the circumference segments overlapping the shape. At each iteration r of the circumference growth, detect each segment and obtain their lengths. Assuming that $\mathscr{L}(r) = \{seg_1(r), ..., seg_l(r)\}$ is the set of segment lengths for a given radius r, compute:

$$avgSizeNsi = \frac{\sum_{r=1}^{r_{max}} \left(\frac{\sum_{for \quad each \quad seg_k \in \mathscr{L}(r)} seg_k}{|\mathscr{L}(r)|} \right)}{r_{max}}. \tag{4.18}$$

7. Variance of the segment length

This measure computes the variance of the sizes of all segments that intersect the shape and is given by Eq. (4.19):

$$varSizeNsi = \sum_{r=1}^{r_{max}} \left(\sum_{for \quad each \quad seg_k \in \mathscr{L}(r)} seg_k - avgSizeNsi \right)^2. \tag{4.19}$$

8. Sum of the first derivative of ovp

This measure computes the energy considering the approximations of the first derivatives for all radius values of the overlapping pixels signature:

$$firstDerNsi = \sum_{r=2}^{r_{max}} \Big(ovp(r) - ovp(r-1) \Big)^2. \tag{4.20}$$

9. Weighted rates

The signature *ovp* is weighted by *nsi* and this measure consists of adding each weighted value as given by Eq. (4.21):

$$nsiOvp = \sum_{r=1}^{r_{max}} \Big(nsi(r) \times ovp(r) \Big)^2. \tag{4.21}$$

10. Fluctuation of nsi

Using the invariant moments of Hu [192] as motivation, Eq. (4.22) describes the fluctuation of the signature *nsi*:

$$flucNsi = \frac{\max_{r=1}^{r_{max}} \Big(nsi(r) \Big) - \min_{r=1}^{r_{max}} \Big(nsi(r) \Big)}{|avgNsi|}. \tag{4.22}$$

11. Fluctuation of ovp

The fluctuation of the signature *ovp* is given by Eq. (4.23):

$$flucOvp = \frac{\max_{r=1}^{r_{max}} \Big(ovp(r) \Big) - \min_{r=1}^{r_{max}} \Big(ovp(r) \Big)}{|avgOvp|}. \tag{4.23}$$

12. Variance of the cumulative sum of ovp

Let *cumsumOvp* be a signature that composes the cumulative sum of *ovp* and *avgcsOvp* the average value of the cumulative signature. The variance of the cumulative signature can indicate whether shapes are compact or round once the cumulative signature of *ovp* presents an uniform and increasing pattern. Eq. (4.24) presents this measure:

$$varcsOvp = \sum_{r=1}^{r_{max}} \Big(cumsumOvp(r) - avgcsOvp \Big)^2. \tag{4.24}$$

13. Shape deficiency

This measure provides the deficiency area of a shape S in regard to the circle with the highest radius value r_{max}. Let $A_{r_{max}}$ be the area delimited by such circle with radius r_{max} and A_{m_1} the area delimited by the circumference for the radius value r_{m_1}. Then, computing the deficiency requires removing A_{m_1} from $A_{r_{max}}$ to produce the area A_{def}, and obtaining:

$$deficiency = \frac{|A_{def} \cap A(S)|}{|A_{def}|}. \tag{4.25}$$

14. Compactness

This parameter provides a measure of the compactness of a shape. Let r_l be the radius value in which $ovp(r_l) < 1$ and $1 \leq r_l \leq r_{max}$. The compactness is obtained as:

$$compact = \frac{r_{max} - r_l}{r_{max}}. \tag{4.26}$$

15. Rate of overlapping pixels in r_{mid}

This rate of overlapping pixels at the specific radius length r_{mid} is given according to Eq. (4.27):

$$ovpMid = ovp(r_{mid}). \tag{4.27}$$

16. Number of intersections in r_{mid}

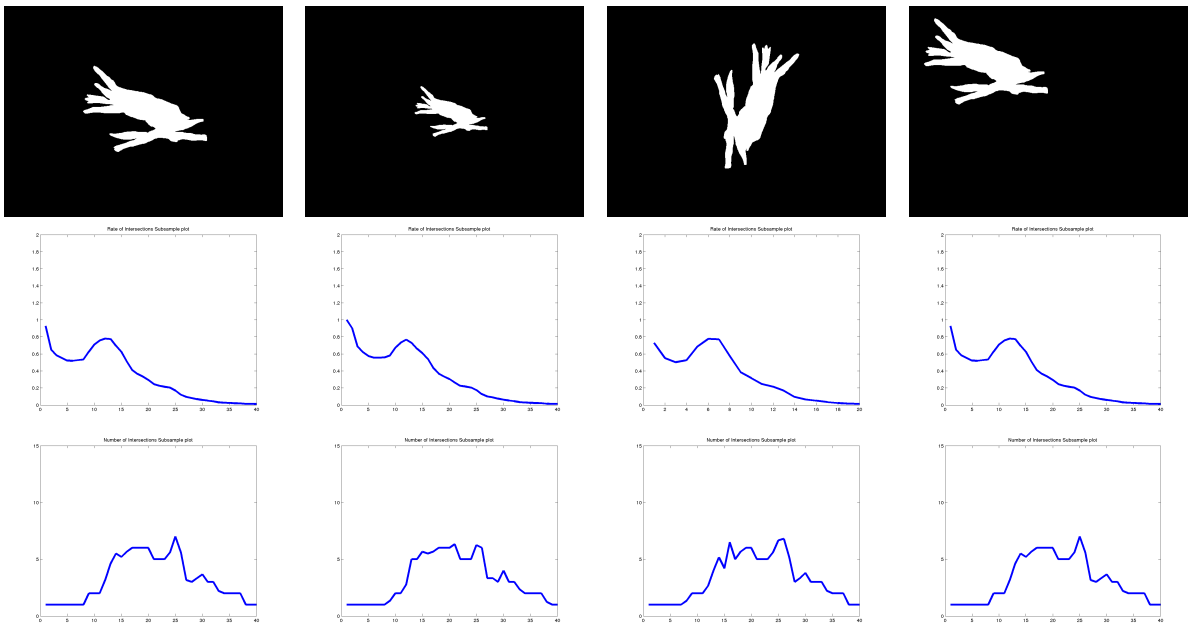
The number of intersections between the circumference of radius size r_{mid} and the shape is obtained by Eq. (4.28):

$$nsiMid = nsi(r_{mid}). \tag{4.28}$$

Figure 60 illustrates a dependence analysis of the proposed SID with respect to the geometric transformations: scale, rotation and translation. For each column of images, the first row depicts the alga shapes, and the second and third rows show the associated signatures *ovp* and *nsi*, respectively. The image in Figure 60(a) presents an original alga shape S_{orig} and its signature, which are taken as reference for comparison with the signatures obtained from the rotated, scaled and translated shape by means of the mean squared error e.

Figure 60(b) presents the scaled shape S_{sca} and its associated signatures, the mean squared errors are $e_{ovp}(S_{orig}, S_{sca}) = 0.12$ and $e_{nsi}(S_{orig}, S_{sca}) = 4.24$. Figure 60(c) depicts in the first row the rotated shape S_{rot}, the *ovp* and *nsi* signatures in the remaining rows. The mean squared errors are $e_{ovp}(S_{orig}, S_{rot}) = 0.34$ and $e_{nsi}(S_{orig}, S_{rot}) = 4.24$. Figure 60(d) shows the

translated shape S_{tra}, its signatures *ovp* and *nsi*, with $e_{ovp}(S_{orig}, S_{tra}) = 0.0$ and $e_{nsi}(S_{orig}, S_{tra}) = 0.0$.



(a) (b) (c) (d)

Figure 60 – Invariance to geometric transformations. Top-down analysis per column: the shape, subsampled ovp signature and subsampled nsi signature: (a) original alga shape; (b) subscaled alga shape; (c) rotated algae; (d) translated algae.

In general, SID presents low variance in regard to the aforementioned geometric transformations. First, it is invariant to translation since it does not depend on the shape position in the image space for detecting its intersections with the circumferences growing from its centroid. It is relatively invariant to scale because the two signatures are subsampled and by the fact that statistical, quantitative and qualitative measures are computed from this simplified representation. Finally, it has low variance in regard to rotation since the strategy for computing the centroid is based on the skeleton, which is a representation invariant to rotation. In a global perspective, SID is capable of preserving the shape signature patterns, which is a desirable characteristic of shape descriptors.

4.4 Proposed green alga descriptors

In the conventional approach to taxonomical identification of Selenastraceae algae, biologists first observe some specific morphological shape properties, e.g. if an alga appear as isolated cells or organized in colonies, the shape geometry (elongated, round or ellyptical) and the presence of certain elements (such as pirenoid or mucilage). Shape features of different natures were selected in order to effectively capture the various types of morphological characteristics and the organization of alga cells.

Bearing this in mind, the strategy attempt to capture and quantify the relevant proper-

ties as statistical measurements that are directly related with such morphological characteristic. Considering the description in Section 2.1.3, the morphological characteristics of the species considered in this research can be summarized below for the properties of isolated alga cells and colonies (when applicable):

- **1.** *Ankistrodesmus densus*
  - Isolated cell: elongated and average to large length
  - Colony: Stellate shape or "Bifurcate"-shaped and average-to-large length
- **2.** *Selenastrum bibraianum*
  - Isolated cell: C-shaped with varying concavity size
  - Colony: Stellate shape or small-to-averate length
- **3.** *Raphidocelis subcapitata*
  - Isolated cell: Small, round, C-shaped and with the presence of a concavity
  - Colony: Small and round shapes
- **4.** *Kirchneriella aperta*
  - Isolated cell: round shapes with a concavity and small size, but bigger than *Raphidocelis subcapitata* isolated cells
  - Colony: enqueued or overlapping round cells
- **5.** *Monoraphidium griffithii*
  - Isolated cell: elongated with long length and sharp
  - Autosporia: Spicullate-shaped or "Bifurcate"-shaped
- **6.** *Monoraphidium contortum*
  - Isolated cell: Small filaments with curved shape
- **7.** *Ankistrodesmus fusiformis*
  - Isolated cell: Elongated, sharp and an average length
  - Colony: Thick cells joined in their middle

Using these properties as the main reference to design a green alga descriptor, the strategy is to manually map the morphological characteristics above to corresponding computational shape features. For instance, the morphological characteristic "elongated" can be described by the shape diameter or by the aspect ratio of its minor and major axis. The roundness of

an alga shape can be measured by the mean curvature and the circular descriptor statistics, and the colony behavior by one of the measures given by SID.

The problem of characterizing green algae is extremely complex, requiring a combination of representative features capable of distinguishing the species. In this work, two alga descriptors are proposed: one uses a combination of basic shape geometric features and the Curvature Scale Space descriptor; the other employs some measures of SID and combined with other basic shape geometric features.

The shape features of these two descriptors are evaluated according to the contribution of each feature by measuring the gain ratio with respect to each alga species. The main goal in employing this strategy is to determine to which extent a particular feature can individually discriminate the instances into the possible species. Such measures are also important to model a customized decision tree classifier for Selenastraceae identification, to select the relevant attributes for the tree nodes. The Gain Ratio measure, which incorporates the Information Gain with an entropy criterion, has been chosen due to its wide applicability in decision tree learning algorithms [129] and for feature selection.

Let A_j be a feature (attribute) from the attributes set $\mathscr{A} = \{A_1, ..., A_m\}$. $\mathbf{X} = \{\mathbf{x_1}, ..., \mathbf{x_n}\}$ is the set of training instances and $\mathscr{A}$ describes the images associated to $\mathbf{X}$. $\mathscr{L}$ is the set of class labels $\{l_1, ..., l_L\}$, i.e., the known Selenastraceae species. The function $value(l_k, j)$ returns the class label for the attribute A_j. The criterion for measuring the degree of impurity of an instance set is the entropy, in which $H(l_k)$ denotes the entropy of class l_k [98]. The information gain for an attribute A_j relative to a class label l_k is defined by Eq. (4.29):

$$IG(\mathscr{L}, A_j) = H(\mathscr{L}) -$$

$$\sum_{v \in A_j} \left(\frac{|\{l_k \in \mathscr{L} | value(l_k, A_j) = v\}|}{|\mathbf{X}|} H(\{l_k \in \mathscr{L} | value(l_k, A_j) = v\}) \right), \tag{4.29}$$

in which $|\mathbf{X}|$ is the cardinality of $\mathbf{X}$. The intrinsic information for partitioning the dataset using A_j is computed considering the entropy of distribution of instances into such partitions, i.e., the contribution required to assign an instance to a partition is given by:

$$InI(\mathscr{L}, A_j) = -\sum_{v \in A_j} \left(\frac{|\{l_k \in \mathscr{L} | value(l_k, A_j) = v\}|}{|\mathbf{X}|} \right) \log \left(\frac{|\{l_k \in \mathscr{L} | value(l_k, A_j) = v\}|}{|\mathbf{X}|} \right). \tag{4.30}$$

The gain ratio is computed as:

$$GainRatio(\mathscr{L}, A_j) = \frac{IG(\mathscr{L}, A_j)}{InI(\mathscr{L}, A_j)}. \tag{4.31}$$

The higher is the $GainRatio(\mathbf{X}, A_j)$, the more important and relevant is the attribute for correctly classifying the dataset. For each descriptor, the gain ratio is provided as a measure of the relevance of the selected shape features.

4.4.1 Basic descriptor

This descriptor contains some basic geometrical measures from the alga shapes and combines them with a complex shape descriptor known as Curvature Scale Space. Particularly, the choice of the Curvature Scale Space (CSS) descriptor is due to its successful applications in a previous work that aims to recognize diatoms using their morphological characteristics [112]. In addition, as some Selenastraceae species can be distinguished using the alga shapes, CSS might describe their main relevant patterns.

The Curvature Scale Space (CSS) [176] is a well-known descriptor that captures the key local shape features by representing the shape boundary curvatures in a scale space which describes the locations of convex (or concave) segments and also detects the degree of convexity (or concavity) of such segments. The scale space representation of a shape is created by tracking the position of inflection points in a shape boundary filtered by low-pass Gaussian filters of variable widths (Eq. 4.32):

$$g(x,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{||x||}{2\sigma^2}\right) \tag{4.32}$$

in which σ is responsible for balancing the smoothness. As the width of the Gaussian filter increases, negligible inflections are removed from the boundary and the shape becomes smoother. The remaining inflection points in the representation are likely to describe relevant object characteristics. Figure 61(a) shows the binary region of interest and Figure 61(b) the original contour subsampled to 200 points. Figures 61(c) and 61(d) present the smoothed contours by setting $\sigma = 0.3$ and $\sigma = 16.0$, respectively.

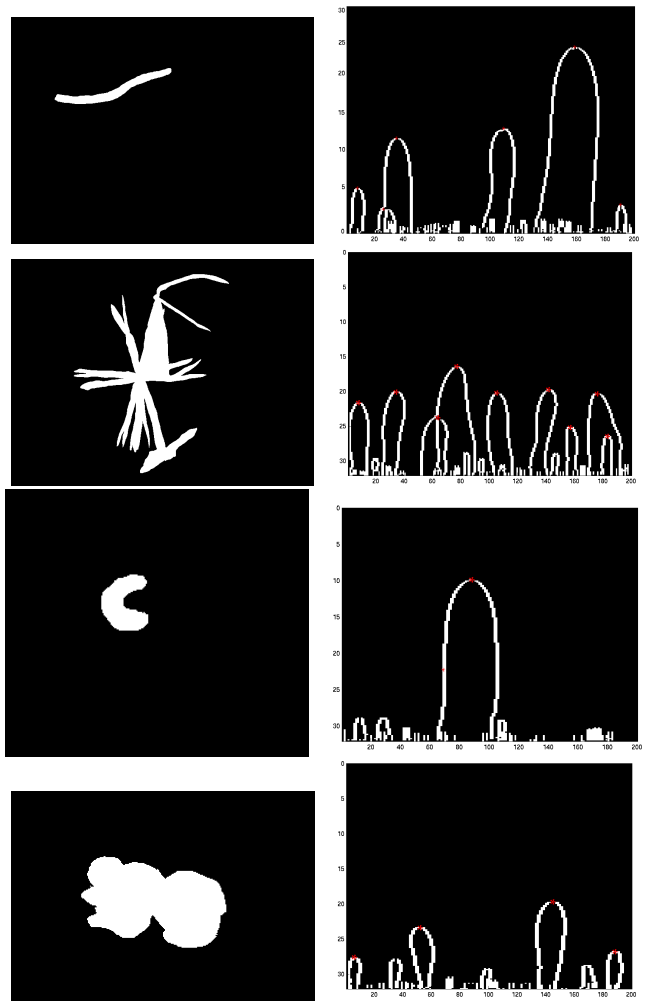
(a) (b) (c) (d)

Figure 61 – Contour smoothing: (a) Binary image; (b) image contour; (c) smoothed contour using $\sigma = 0.3$; (d) smoothed contour using $\sigma = 16.0$;

This multi-scale smoothing process produces a map depicting an interval tree formed by several inflection points. Figures 62 presents four alga shapes and their respective CSS maps. Figure 62(b) show the CSS maps of the alga shapes depicted in the rows of Figure 62(a), in which the red points are the maxima. The $x-$axis presents the arc length of the alga contour.

The $y-$axis refers to the width of the Gaussian low-pass filtering in the contour. The maxima points are determined by reading the CSS map pixel-by-pixel and finding the peaks by means of a 5×5 kernel. More details of how to compute the CSS map from a binary image can be found elsewhere [193].

It is noticeable that maps of groups of cells in the second row have more points of maxima than maps of single alga in the first row. The bottom rows depict two cases of rounded algae, in which the maxima in the third row express the concavity of the shape and the low magnitude of maxima in the fourth row denotes its compactness. Thus, this descriptor can be useful for distinguishing between different alga species, in particular their colonial behavior, for which the basic geometric features are not appropriate.



(a) (b)

Figure 62 – Curvature Scale Space maps (one case per row): (a) alga shapes; (b) respective CSS maps.

Two CSS descriptors are compared by means of a matching algorithm, which compares the associated sets of maxima points and assigns a matching cost value as a measure of dissimilarity. The matching process takes into account a circular shift and a translation along the respective x-axis of the maxima, to obtain invariance to rotation. Considering a query image, the idea is to compare its extracted maxima with the maxima sets of the entire image set. When

comparing two maxima sets with the distance function, the lowest cost among the possible combinations expresses the best possible match. After comparing the query image to the entire image set using the maxima points, the output of the matching process is a rank with the images in the set displayed in increasing order of dissimilarity. A conventional strategy is to assign to the query image the most frequent label in the K most similar images.

Table 2 summarizes the potential association of each morphological characteristic to a shape feature. Note that the behavior of alga cells is described by the CSS descriptor, which provides a set of maxima, while the remaining features are single measures describing a shape property. Moreover, Table 2 presents the average gain ratio (Eq. (4.31)) computed for each basic geometric feature in relation to the class labels on a set of 56 green alga images. This set contains 8 sample images from each of 7 species, as described in Section 5.1.2. In images with multiple algae the cell with the largest area is selected as its representative. Area is the most discriminative shape feature for this particular green alga set. On the other hand, average bending energy shows low discrimination capability, suggesting it might be more useful to distinguish subsets of species.

Table 2 – Basic descriptor.

id	Morphological characteristic	Computational shape feature	Gain ratio	Standard deviation
1	Recognition of stellate Colony/Autosporia	CSS / Area	0.862	0.195
2	Recognition of round cells	Convexity	0.581	0.318
3	Recognition of elongated isolated cell	Rectangularity	0.784	0.285
4	Presence of concavity (V-Shaped)	Solidity	0.585	0.247
5	Concavity size	Convex defficiency	0.585	0.247
6	Colony size	Perimeter	0.738	0.186
7	Length isolated cell	Diameter	0.826	0.268
8	Presence of mucilage	Eccentricity	0.803	0.252
9	Thickness	Th	0.54	0.251
10	Cell curvature	Average bending energy	0.12	0.349

Table 2 also presents the standard deviation values of each shape feature, considering they are normalized to $[0,1]$. Such values indicate considerable variability within the shape features, with the higher value associated with the feature ABE on Eq. (4.3). This scenario may indicate that basic geometric features are not sufficiently discriminative for alga species, although some of them, such as diameter and area, can be combined with other descriptors.

The next step requires modeling a suitable dissimilarity measure to compare the feature vectors given this descriptor. The computed features are of distinct natures, making it impossible to measure vector dissimilarity with a simple metric function.

A metric function can be used to compute the distance among the basic geometric features, while the dissimilarity measure to compare two CSS signatures is given by the CSS matching process. Here, the distance function $d_{GF}(f_1, f_2)$ denotes the distance between two sets of basic shape features f_1 and f_2 and the matching cost for comparing two CSS descriptors is denoted by $\mathscr{M}$. Finally, the dissimilarity measure that compares two green alga images I_1 and I_2, represented by their respective feature vectors $\{f_1, CSS_1\}$ and $\{f_2, CSS_2\}$, is obtained with

Eq. (4.33):

$$d_{BD}(I_1, I_2) = d_{GF}(f_1, f_2) + \mathscr{M}(CSS_1, CSS_2) \tag{4.33}$$

This descriptor may be employed with a K-Nearest Neighbor classifier as the only way to compare CSS descriptors is by means of a matching process. Finally, after obtaining pairwise image comparisons $d_{BD}(I_i, I_j)$ for a set of n images $\{I_1, ..., I_n\}$, a $n \times n$ distance matrix $\mathscr{D}$ is obtained, in which $\mathscr{D}(i, j) \approx d(I_i, I_j)$, for $i = 1, ..., n$ and $j = 1, ..., n$ can be used as input to a classifier or to visualization techniques.

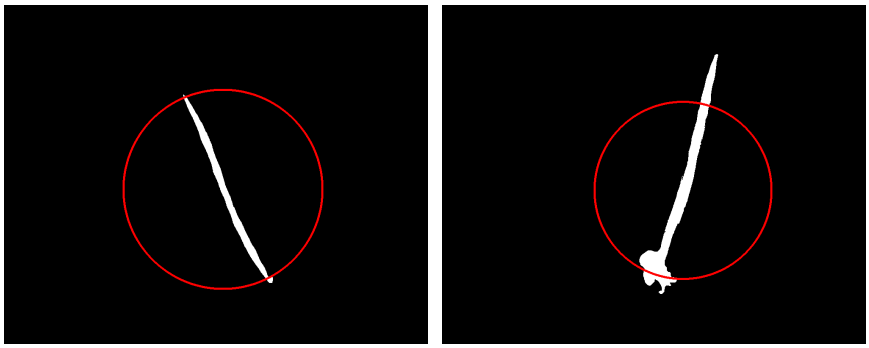
As described in Section 4.5, and also corroborated by the features gain ratios and variances, this descriptor is not sufficiently discriminative in relation to species. The combination of basic geometric features with the CSS map and the proposed dissimilarity measure have not succeeded in characterizing the green alga species, as further discussed on Chapter 5. As a result, this work proposes an additional green alga descriptor that combines some basic shape geometric features and some SID measures.

4.4.2 SID-based green alga descriptor

A new green alga descriptor is introduced as a combination of basic shape features with other derived measures from the SID descriptor. This combination can potentialize the discrimination of alga species and the description of their morphological characteristics.

The process of mapping the morphological characteristics of green algae to SID features consisted of visual analysis of the signatures patterns and their derived metrics. For instance, elongated instances present lower values of compactness, hence this feature is useful for distinguishing between elongated and rounded cells. Smaller alga cells might present shorter signatures and rounded cells can be characterized by SID's compactness measure. Furthermore, the feature *avgSizeNsi* captures the segments' lengths, useful for describing alga shapes with concavities.

The mucilage in elongated cells can be detected by verifying the two corners of the alga body. Figures 63(a) and 63(b) depict two *Monoraphidium griffithii* algae, in which the alga in Figure 63(b) presents mucilage. Using SID, the algorithm to detect the presence of mucilage requires that alga be elongated. It is possible to detect this pattern in an alga shape by checking whether $avgNsi < 3$. Then, the circumference is grown from the shape's centroid until obtaining no more than one intersection with the shape. At this moment, the radius value r_{one} is stored and the radius mucilage $r_{muc} = r_{one} - 5$ is set to obtain the circumference overlapping the mucilage region and the opposite corner region of the algae, as shown in Figure 63(b). The length of each segment identified by the circumference overlapping the shape is obtained and denoted as n_{muc} and n_{cor}, respectively. Mucilage is present in an elongated alga cell if the condition $n_{muc} > 2n_{cor}$ is satisfied. In Figure 63(b), $n_{muc} = 49$ and $n_{cor} = 15$, so mucilage is recognized.



(a) (b)

Figure 63 – *Monoraphidium griffithii* and the SID descriptor: (a) without mucilage; (b) the presence of mucilage.

An experiment was conducted to evaluate the contribution of the SID measures to class labels on the same aforementioned set of 56 green alga images, again taking the cell with the largest area as the image representative. The experiment consists of computing the average gain ratios according to Eq. (4.31) for each feature. After that, SID measures are sorted considering their gain ratios and Table 3 presents the results, in which the non-contributing features are ignored.

Table 3 – Gain ratios of SID features in relation to the classes on a set of 56 green alga shapes.

ID	SID feature	Gain ratio
14	*compact*	0.88
11	*flucOvp*	0.816
15	*ovpMid*	0.775
4	*avgOvp*	0.767
6	*avgSizeNsi*	0.673
9	*nsiOvp*	0.567
13	*deficiency*	0.51
10	*firstDerNsi*	0.488
7	*varSizeNsi*	0.463
3	*varNsi*	0.463

Table 4 summarizes the SID-based features that are potentially assigned to describe each morphological green alga characteristic and their standard deviations, bas ed on the previous. The gain ratio of each feature has been computed on the same aforementioned shape set of 56 alga images. In general, SID measures obtain better gain ratios when compared with the features of the Basic descriptor. However, the variability within the features is still high, suggesting that this undesirable aspect is a pattern of the green alga data. Therefore, the strategy of selecting the shape features qualitatively can improve the correct classification rates of green alga species and support the modeling of a specialized decision tree for classification, as detailed in Chapter 5.

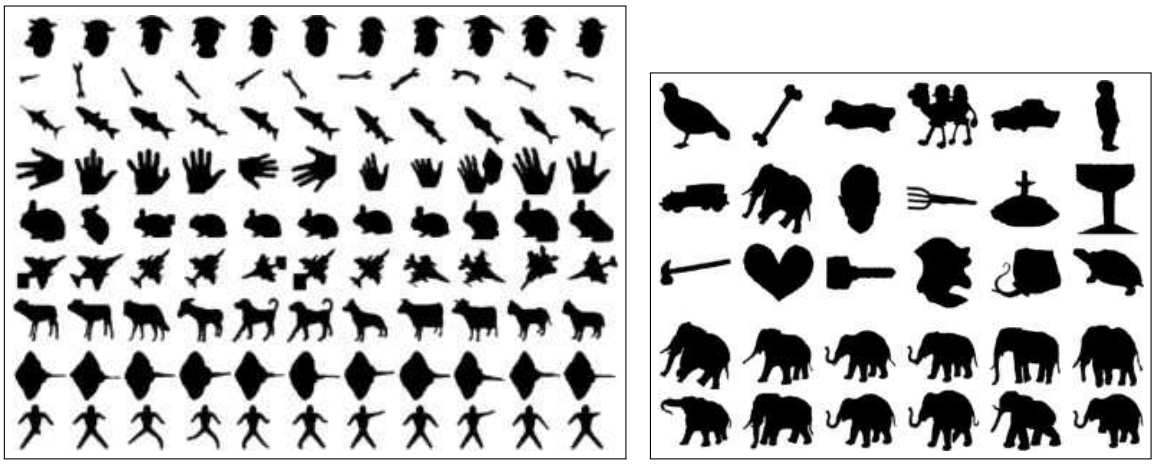
Table 4 – SID-based green alga descriptor.

id	Morphological characteristic	Shape feature	Gain ratio	Standard deviation
1	Recognition of stellate colony/autosporia	$flucOvp$	0.816	0.195
2	Recognition of round cells	$compact$	0.88	0.218
3	Recognition of elongated isolated cell	$deficiency$	0.51	0.285
4	Presence of concavity (V-Shaped)	$avgOvp$	0.767	0.247
5	Concavity size	$avgSizeNsi$	0.673	0.247
6	Colony size	Area	0.862	0.195
7	Length isolated cell	Diameter	0.826	0.268
8	Presence of mucilage	Derived methodology	0.23	0.252
9	Thickness	Th	0.54	0.251
10	Cell curvature	$firstDerNsi$	0.448	0.212

4.5 Experimental results

Although SID has been proposed to describe morphological characteristics of green alga shapes, some experiments were performed aiming to evaluate its generalization capability to shapes of different natures. Particularly, the goal of these experiments is to assess the effectiveness and performance of SID on classification tasks on diverse kinds and compare it with some selected shape descriptors from the literature. For that purpose, this work follows the evaluation strategies adopted in shape-based image retrieval systems, also known as *Content-Based Image Retrieval* (CBIR) [194]. In addition, a comparative analysis of computational complexity with existing shape descriptors is provided.

Two popular sets for evaluating shape descriptors or matching algorithms are the Kimia-99 and Kimia-216 [195]. The first one is a shape set with 99 binary images depicting nine categories of real-world objects. Kimia-99 is quite challenging because some shapes contain occlusions, missing parts, and articulations. Figure 64(a) illustrates the shapes, in which each row corresponds to a shape category. Kimia-216 [195] consists of 216 shapes, grouped in 18 classes with 12 samples of each class. Some assorted examples of these shapes are shown in Figure 64(b) and most shapes belong to MPEG-7 Part-B set [196].



(a) (b)

Figure 64 – Samples of each shape set: (a) Kimia-99; (b) Kimia-216.

The measure to evaluate and compare the shape descriptors is the retrieval accuracy rate. It can be computed by the Bull's eye test [197], which measures the correct matches among the top retrievals regarding the number of similar shapes. This strategy consists of using each shape as a test query, retrieving shapes from the same set and ranking them according to the similarity to the query. Shapes in the same category are expected to be top-ranked while dissimilar shapes are not.

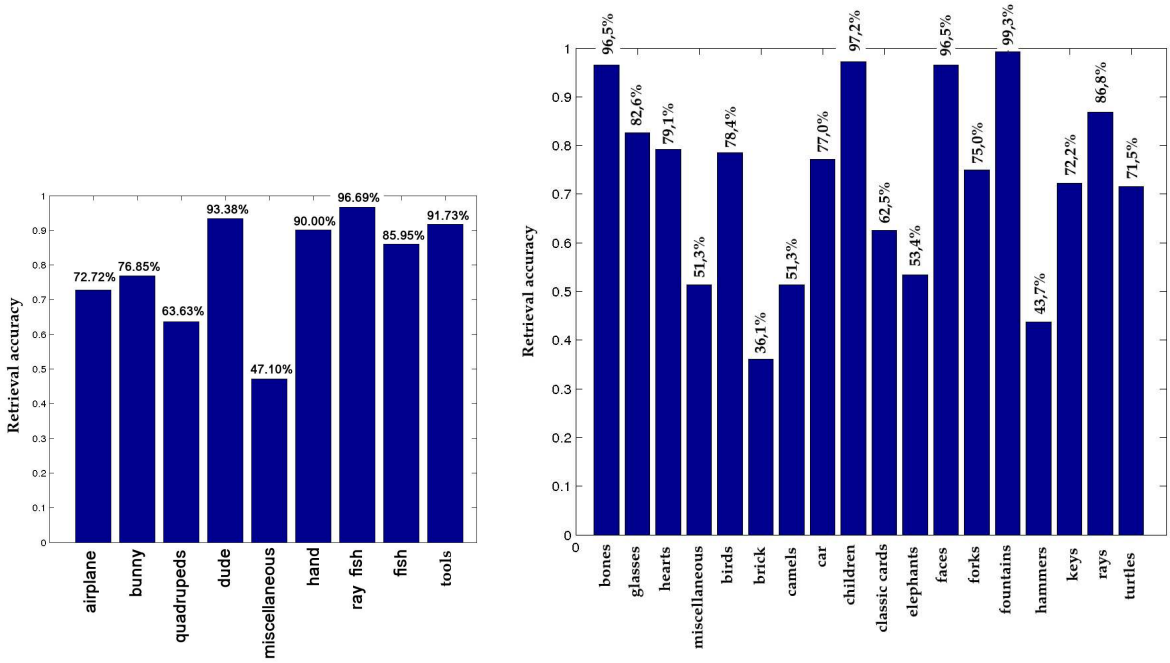
The Bull's eye test, for every image in a set of N images grouped into N_C classes, the number of shapes per class $N_S = \frac{N}{N_C}$ and the top $2N_S$ most similar shapes are retrieved. At most half of the retrieved shapes are correct hits. The retrieval accuracy is measured as the rate of the number of correct hits (considering all queries) and the highest possible number of correct hits. On Kimia-99, $N_S = 11$, meaning that in the top 22 retrieved shapes, the best possible result is $N_S = 11$ correct matches for each shape query and the highest number of correct hits is 99. On Kimia-216, $N_C = 18$ and among the top 36 retrieved shapes, the maximum number of correct matches is $N_S = 12$ for each query.

Table 5 reports the classification performances on Kimia-99 for SID and five state-of-the-art shape descriptors. The table shows a similarity retrieval rank, in which one shape is selected as the query and then matched to the remaining shape contours. The number of correct matches within the top 10 best matches is counted by verifying how many shapes belong to the same class as the query. The final column shows the correct retrieval rate for the top ranked 10, in which SID achieved the second best rate, despite its lower dimensionality. The IDSC [198] descriptor achieved the highest correct classification rate, but it is important to emphasize that IDSC is time consuming because is requires a matching process for measuring similarity between shapes.

Table 5 – Kimia-99: Retrieved similarity rank.

Method	1^{st}	2^{nd}	3^{rd}	4^{th}	5^{th}	6^{th}	7^{th}	8^{th}	9^{th}	10^{th}	Total
SID	99	95	94	87	92	87	81	81	70	62	85.40%
Zernike Moments [199]	99	89	83	82	83	81	77	76	71	70	63.99%
Fourier Descriptor [200]	95	93	87	86	78	70	65	61	55	41	73.83%
Curve Normalization [201]	97	86	87	75	76	70	55	59	46	44	70.0%
Shape Context [202]	97	91	88	85	84	77	75	66	56	37	76.36%
IDSC [198]	99	99	99	98	98	97	97	98	94	79	96.67%

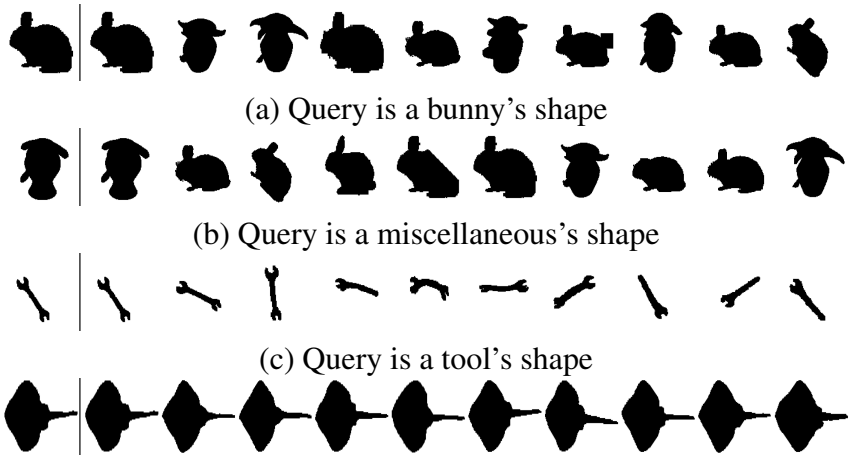
Figure 65 presents the accuracy retrieval rates for each shape category on Kimia-99 and Kimia-216 sets. These measures were computed considering the top-10 and top-12 most similar retrieved shapes using the scheme employed in the Bull's eye test for Kimia-99 and Kimia-216, respectively. In both shape sets, some classes achieve retrieval accuracy rates above 90% which confirms the applicability of SID to describe shapes other than green alga shapes. Some poor retrieval accuracy rates are obtained for categories in which their overall global aspect is similar and local properties are of fundamental importance to produce discrimination. For a more detailed analysis, some extra similarity retrieval ranks are shown for both shape sets.



(a) (b)

Figure 65 – Accuracy retrieval rates per class: (a) Kimia-99; (b) Kimia-216.

Figure 66 presents retrieval similarity ranks for four shapes from different categories on Kimia-99. The queries are shown on the left and the top-10 retrieved shapes are ordered according to their similarity to the query shape. The query shapes in the first and second rows are similar according to SID because their similarity ranks contain retrieved shapes from both categories, but they actually belong to different shape categories. The third and fourth rows present queries of two different shape categories, in which their similarity ranks yielded the highest accuracy retrieval rates.



(a) Query is a bunny's shape

(b) Query is a miscellaneous's shape

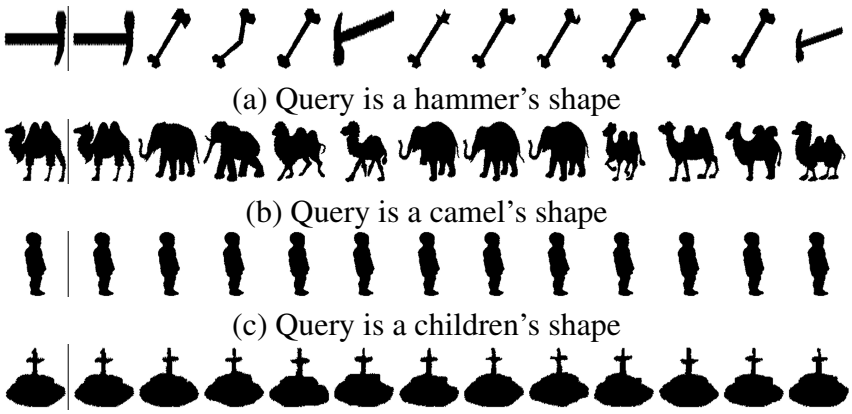
(c) Query is a tool's shape

(d) Query is a ray fish's shape

Figure 66 – Top-10 retrieved shapes of some query shapes from Kimia-99 using the SID descriptor.

Some retrieval similarity ranks for query shapes in Kimia-216 are illustrated in Figure 67. The queries are from four different shape categories and the top-12 most similar shapes

are retrieved in each case. In the first row, showing a query from the *Hammer* category, several shapes from the *Bone* category were retrieved. The retrieval accuracy rates for the *Bone* and *Hammer* categories are $96,5\%$ and $43,7\%$, respectively. In the latter, hammer is mostly recognized as bone. The second row refers to a query from the *Camel* category and the similarity rank contains some shapes from the *Elephant* category. The retrieval accuracy rates for the *Camel* and *Elephant* categories are $51,3\%$ and $53,4\%$, respectively. The third and fourth rows present queries of shape categories, *Children* and *Fountain*, respectively and their similarity ranks reflect the higher accuracy rates above 95%.



(a) Query is a hammer's shape

(b) Query is a camel's shape

(c) Query is a children's shape

(d) Query is a fountain's shape

Figure 67 – Top-12 retrieved shapes of some query shapes from Kimia-216 using the SID descriptor.

A comparison of the computational complexities of SID and some selected shape descriptors from the literature is provided. Table 6 compares the dimensionality of the feature vectors resulting from the different descriptors and the complexities for computing the features and for measuring the pairwise shape dissimilarities. The Big-O notation is used to express the computational time cost as a function of the input length n. The analysis allows to infer that SID scores better, in general, than the methods from the literature regarding dimensionality and complexity of dissimilarity computation.

Table 6 – Dimensionality and computational time complexity.

Shape descriptor	Number of attributes	Complexity
Beam Angle Statistics [203]	180	$O(n^3)$
Triangle Area Representation [204]	180	$O(n^3)$
Zernike Moments [199]	36	$O(nlogn)$
Shape Context [202]	45	$O(n^2)$
Fourier descriptors [199]	126	$O(n^2)$
SID	16	$O(n^2)$

Table 6 indicates that SID is a useful shape descriptor that allies reduced computational cost with good retrieval accuracy, as compared with reference methods. SID is a descriptor that characterizes shapes globally because it computes measures from the entire signatures. In this sense, it is indicated for describing shape sets in situations where the computational cost is a

concern, while preserving the ability to retrieve or classify shapes accurately.

Nonetheless, SID has a major limitation of not capturing effectively minor details or local patterns in shapes, such as the sizes of small concavities and their positioning along the shape. Such information in shapes is not properly well captured since SID represents the local properties of a shape, as well as such details might be diluted along the signature. Moreover, SID showed to be variant to the presence of shape deformations, since curved shapes with shallow concavities have high variations in their signatures when compared to signatures of normal shapes in a same category.

4.6 Final considerations

This chapter presented the proposed descriptor for green microalga images and a general shape descriptor named Segment Intersection Descriptor (SID). The green alga feature descriptor can be summarized as a set of basic shape geometric measures, such as area, perimeter and diameter, plus other features describing complex shape patterns, such as colonies, roundness, elongation and the presence of mucilage. As most attributes are numeric, computing the distance among feature vectors can be performed using a distance function.

SID is a set of quantitative and statistical measures computed from two shape signatures. These measures can describe several shape properties concerning geometry and topology and are invariant to scale, translation and relatively to rotation. The experiments reported some promising results of applying SID on challenging and noisy benchmark shape sets. Besides capturing a global perspective of the shape and having a low computational cost, this descriptor retains good accuracy in classification using the Euclidean distance to compute dissimilarity.

In Chapter 5, the classification of green microalga images is addressed by means of two approaches: one automatic and the other incremental user-guided with the support of tree-based visualizations.

CHAPTER

5

CLASSIFICATION OF GREEN MICROALGA IMAGES

The previous chapter described the shape descriptors introduced for the characterization of green microalgae in digital images. From the feature vectors and associated dissimilarity measures it is possible to learn a classification model from training images and then predict the class of unknown species in new images.

Selecting a classification model is a complex decision, as a wide variety of choices could be appropriate to the problem characteristics. Once a classifier is selected, other concerns arise creating appropriate training sets and adjusting classifier parameters. These factors can be difficult to handle and directly influence classification results.

A visual classification approach can minimize such shortcomings which are inherent to supervised classifiers. Visualizations that include the user into the process of discovering knowledge from data can improve the quality and interpretability of classification models. Particularly, user-driven classification of green algae is interesting once the similarity relations between descriptors can be represented in NJ-tree visualizations. Moreover, biologists can rely on their knowledge to interpret such visualizations in favour of the learning process, e.g., to select representative instances for training sets.

This chapter is organized as follows: Section 5.1 introduces a customized decision tree classifier for the alga shapes, and presents results on automatic classification of green microalga images, obtained with the proposed decision tree and with traditional classifiers. Section 5.2 describes a classification methodology that relies on a user-driven exploration process in which a biologist interacts with visualizations in an iterative classification procedure. A case study is presented describing the biologist performing the incremental classification and the results obtained. Section 5.3 presents the final considerations and summarizes the discussions of this chapter.

5.1 Automatic classification

Image classification is a challenging task, as no classification technique works well in all situations. Several aspects affect performance, such as the mathematical formulation of the classification models, the feature space and the dissimilarity measure when applicable. Therefore, selecting a suitable classification model requires solid knowledge about the possible scenarios and data characteristics [205].

In this work, the initial choice of classifiers considered mainly their previous application in previous related work, as discussed in Section 2.3, in addition to the characteristics of the feature space. The green alga data probably presents a nonlinear nature, considering the number of attributes of feature vectors and the fact that some distinct species share common morphological characteristics. Applying a linear classification model for a nonlinear training data can lead to a poor performance due to the low discriminability between instances of different classes. Hence, classifiers which employ optimization processes in their learning algorithms are expected to work well in this scenario [206]. Thus, the Support Vector Machine (SVM) and the Artificial Neural Network (ANN) have been selected and also the K-Nearest Neighbors, as the latter adopts a learning approach that is intuitive even to non-experts. The settings and performance of these classifiers are described in Section 5.1.2.

The manual approach based on identification keys conducted by biologists to identify Selenastraceae species follows a process that resembles a decision tree. Thus, the ID3 algorithm has also been considered, as its formulation underlies a few decision trees in the literature [129]. Although the ID3 algorithm is employed to automatically build a decision tree model considering a measure of the quality of shape features associated to morphological characteristics, it yielded poor performances on both proposed green alga descriptors, as described in Section 5.1.2. This motivated the creation of a strategy in which biologists collaborated to model a customized decision tree classifier for the Selenastraceae species.

5.1.1 Customized decision tree

In the standard approach for identifying Selenastraceae algae, biologists first observe certain specific morphological characteristics and properties, e.g. if algae appear as isolated cells or organized in colonies, their shape geometry (elongated or rounded) and the presence of certain elements (such as mucilage). The proposed strategy is to build a decision tree classifier that somehow learns the morphological characteristics of Selenastraceae algae to predict their species.

Considering the set of morphological characteristics, the learning approach splits the original multi-class classification problem into several minor binary classification problems. Each alga species is characterized by a combination of morphological characteristics, e.g., an isolated cell of *Monoraphidium griffithii* is elongated, long, thick and with mucilage. In a decision tree, the path resulting from the decisions taken from the root node until reaching the

associated leaf node represents a species. Figure 68 illustrates the path obtained when classifying an instance of *Monoraphidium griffithii*, where each node registers the properties of a corresponding morphological attribute.

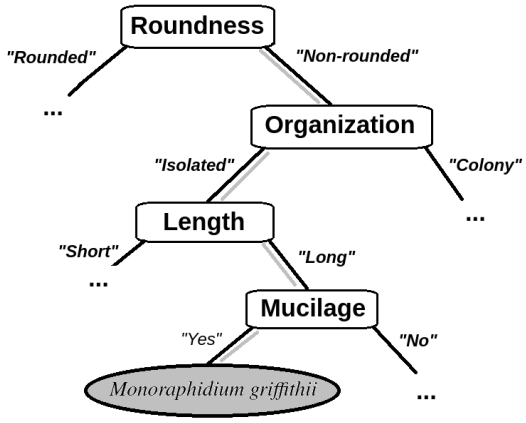
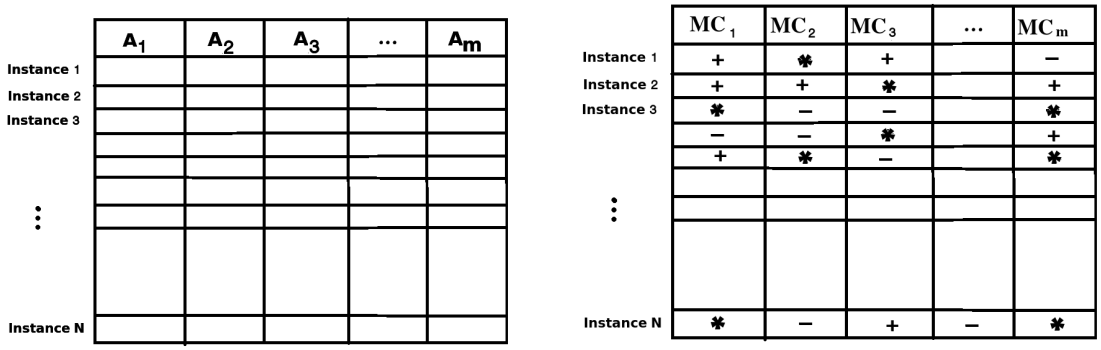


Figure 68 – Path in a decision tree: the gray lines in branches indicate the decisions taken to reach the leaf node from the root.

The learning process of a customized decision tree requires sorting the image attributes (shape features) $\mathscr{A} = \{A_1, ..., A_m\}$ according to their sequence of insertion into the tree. In other words, the first attribute should model the test condition of the root node, the second and third attributes should model the test conditions of the root's left and right childs and so on. For each tree node, its test condition is applied on the attribute value to classify instances relative to an associated morphological property. For instance, a tree node associated with attribute 'roundedness' outputs a binary classification labeling the alga instance as rounded or not-rounded.

The learning algorithm relies on a table that stores the morphological characteristics of the instances on a dataset as a binary model. The role of this auxiliary table is to flag whether and how a morphological characteristic applies to a specific instance by means of values "+1", "-1" or null ('$*$'), meaning that it belongs, respectively, to the positive class, to the negative class or that this morphological characteristic is not applicable. For instance, for a morphological characteristic "roundedness", the positive class indicates rounded shapes and the negative class otherwise. The non-applicable values indicate that this characteristic should be ignored when evaluating the node's test conditions. For instance, mucilage is not present in rounded cells, so it is a characteristic not applicable in this case, but in other cases it can further differentiate between long and thick alga cells, such as instances of *Monoraphidium griffithii* - the positive class (mucilage can be present) - and of *Ankistrodesmus densus* - the negative class (mucilage is not present). Figure 69 illustrates the tables of attributes and morphological characteristics, in which Figure 69(a) refers to the attributes representing the shape features, Figure 69(b) shows the morphological characteristics, and there is a one-to-one mapping between the $j-th$ attribute and the $j-th$ morphological characteristic.

The learning algorithm begins modeling the root node considering the first attribute A_1



(a) (b)

Figure 69 – Proposed decision tree: (a) the table of attributes; (b) the table of morphological characteristics.

and its corresponding morphological characteristic MC_1. For the pair (A_1, MC_1), the average values of A_1 μ_1^+ and μ_1^- are computed for the subsets of "+" and "-" signed-instances in MC_1. The threshold value μ_1 modeling the root node test condition is set as $\frac{\mu_1^+ + \mu_1^-}{2}$. The next pair (A_2, MC_2) models the left-child node, so the decision threshold μ_2 is computed similarly. This process proceeds modeling the root's right child node using pair (A_3, MC_3), and then successively for the remaining attributes until the final attribute. Figure 70 shows a root node and decision rule, in which μ_1 refers to the threshold value computed as described.

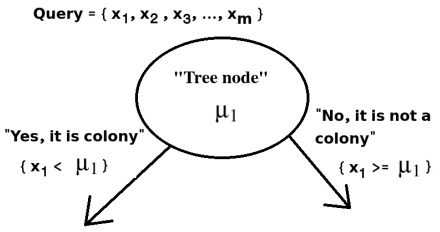


Figure 70 – A node and decision rule in the proposed decision tree.

Figure 71 depicts a node designed to distinguish colonies from isolated cells and its right and left childs. The left-child node handles colony shapes only and decides whether their shape is rounded, the right-child applies a similar test to the single alga case. If the node condition is satisfied ("Yes" situations) it is assumed the left child is the next node to be tested, otherwise ("No" situations) it will be the right-child.

The decision tree structure, i.e., the sequence for inserting nodes into the tree, has been modeled after inspecting a training set with the biologists, and it is depicted in Figure 72. The root node (1) partitions the training set into subsets of rounded or elongated cells, because this morphological property is the most intuitive when analyzing alga shapes. The left-child node (2) differentiates into small and large algae, so it is expected to differentiate both *Raphidocelis subcapitata* and *Kirchneriella aperta* from *Selenastrum bibraianum*. It is worth noting that most

isolated cells of *Selenastrum bibraianum* are characterized as a rounded algae by the shape features in both descriptors. The right-child node (3) aims at distinguishing between individual algae and colonies with elongated cells. *Selenastrum bibraianum* colonies are not rounded, so the node that identifies such cells is placed in the branch of elongated cells with stellate geometry.

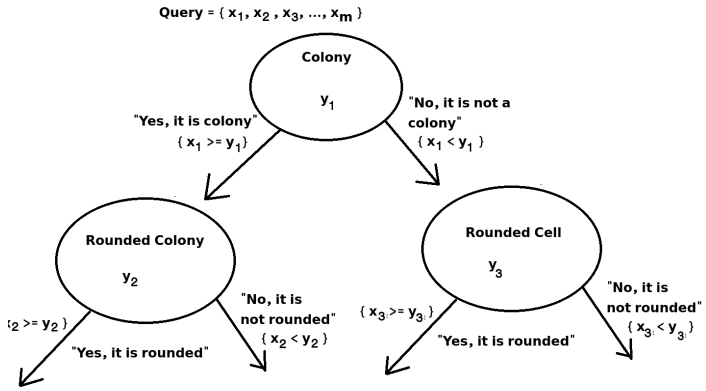


Figure 71 – The root, left-child and right-child nodes of the proposed decision tree after the learning process.

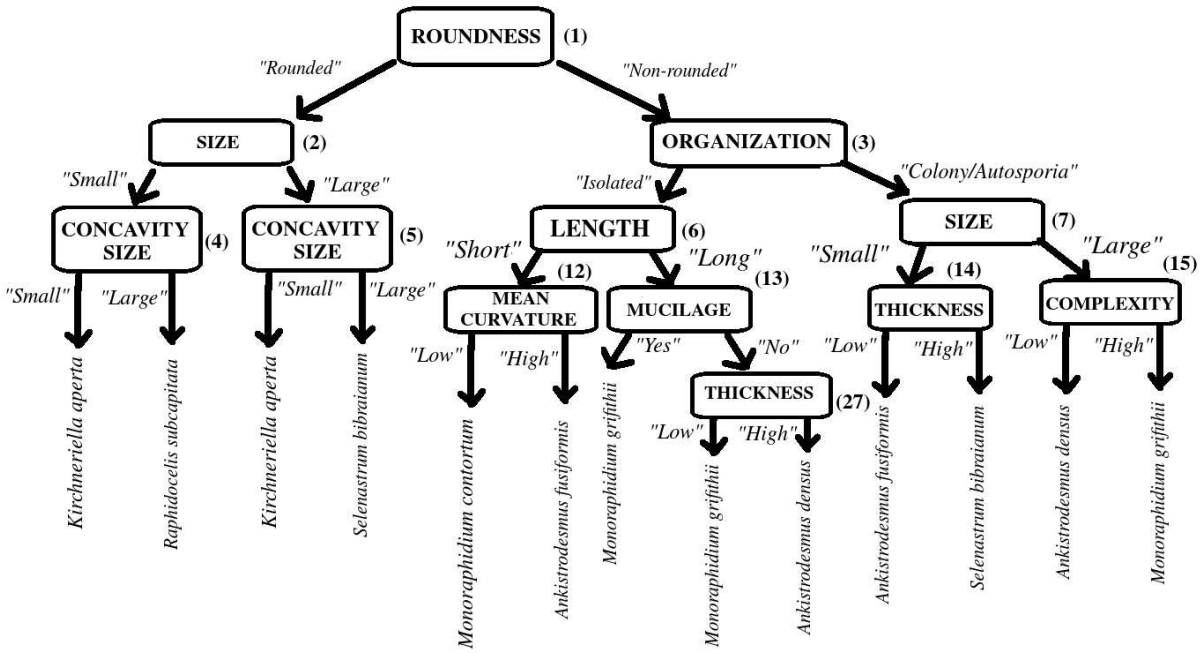


Figure 72 – The customized decision tree.

At the second level of the decision tree, node (4) partitions the rounded and small algae of *Raphidocelis subcapitata* and *Kirchneriella aperta* based on their concavity sizes. In general, isolated cells of *Kirchneriella aperta* present a tight concavity proportional to their area, while *Raphidocelis subcapitata* cells present a significative concave area. Node (5) distinguishes between the isolated cells of *Selenastrum bibraianum* and the colonies of *Kirchneriella aperta*, which present a wide variety of shapes, such as multiple cells with compact rounded shapes or

several concatenated cells with no precise geometry. The children leaves of node (5) identify the two possible species after combining the three shape features (roundedness, size and concavity size) to describe their morphology.

Nodes (6) and (7) attempt to distinguish between short/small and large isolated cells and colonies, respectively. In the first case, species with short cells in length are *Monoraphidium contortum* and *Ankistrodesmus fusiformis*, while *Monoraphidium grifithii* and *Ankistrodesmus densus* have long cells. In the second case, the small colonies are of *Selenastrum bibraianum* or *Ankistrodesmus fusiformis*, whereas *Monoraphidium griffithii* and *Ankistrodesmus densus* appear in larger colonies. Node (12) differentiates between the curved and thin cells of *Monoraphidium contortum* and the straight pattern of *Ankistrodesmus fusiformis* shapes. Node (13) identifies the presence of mucilage, possible in *Monoraphidium griffithii*. On the other hand, as *Monoraphidium griffithii* may not present mucilage, *Monoraphidium griffithii* and *Ankistrodesmus densus* are distinguished based on their thickness in node (27), as the former usually has thiner cells than the latter.

Considering the non-rounded shapes and shapes representing colonies, tree node (14) aims at differentiating species of *Selenastrum bibraianum* and *Ankistrodesmus fusiformis*, in which the former has more compact shapes compared to the thin cells of the latter. Node (15) distinguishes *Monoraphidium griffithii* from *Ankistrodesmus densus*, in which the first presents thiner and longer cells when cells group for autosporia, while in the second the cells join in their respective centers.

The table of morphological characteristics resulted from observing with the biologists the shape patterns and geometries in the training set, and it is fundamental to guide the decision tree learning process. Therefore, the decision tree classifier will perform poorly it its information is inconsistent with the training set. However, observing and annotating shape properties is prohibitive on very large training sets. Future work should consider possible approaches to creating such a table automatically, or alternatively adopting a semi-supervised learning approach.

The effectiveness of this decision tree classifier and other automatic classifiers considered is discussed next.

5.1.2 Experimental results

This section presents experiments performed to evaluate the performance of the classifiers in this problem. They have been performed in a *Intel(R) Core(TM) i7 2.40GHz* using the classifier implementations, evaluation strategies and validation methods available in Weka 3.9 [207]. For the customized decision tree, the tests have been conducted separately in a Java implementation under the same conditions.

In this classification problem, each species is handled as a target class and the classifier assigns to each image a label associated to an alga species. Green alga images may depict multiple alga cells, but they are typically from a single species. For images including multiple algae, the one with the higher area is picked as its representative for classification. An alterna-

tive strategy could rely on generating bounding boxes for each alga cell and then performing classification separately, as described previously in the literature [7] [9]. However, such a strategy would be inconvenient for the visual classification process, which considers images rather than shapes.

Image sets

Experiments were conducted on two different alga image sets to evaluate the performance of distinct classification algorithms and the proposed green alga descriptors. The first image set, named GA-56, consists of ground-truth images, i.e., alga shapes were manually segmented by the biologists. This shape set has 56 images of algae from 7 species, with 8 samples per species. The second shape set, denoted as GA-123, consists of 123 images segmented with the region growing methodology described in Chapter 3. GA-123 is not balanced, i.e., each species has a different number of sample images. Further details are given in Table 7:

Table 7 – Green alga image sets.

Label (id)	Samples in GA-56	Samples in GA-123	Species
1	8	18	*Ankistrodesmus densus*
2	8	36	*Selenastrum bibraianum*
3	8	17	*Raphidocelis subcapitata*
4	8	15	*Kirchneriella aperta*
5	8	17	*Monoraphidium griffithii*
6	8	12	*Monoraphidium contortum*
7	8	8	*Ankistrodesmus fusiformis*
TOTAL	56	123	-

Classifier settings

The first classifier is an Artificial Neural Network (ANN) based on a feed forward multilayer perceptron (MLP) that uses a backpropagation algorithm for the learning process. A sigmoid function is employed as the activation function and each neuron transfer function is modeled by the hyperbolic tangent. The ANN architecture consists of three layers. The input layer is defined by 10 nodes (determined by the size of both proposed green alga descriptors). The output layer requires 4 nodes for representing the class labels (ids $\{0,1,2,3,4,5,6\}$ in a binary numeral system). The parameters *momentum value* and the *learning rate* used in the learning process and the backpropagation algorithm were set respectively to 0.2 and 0.3 after some experimentation (these values yielded the best correct classification rates from other tested values). The number of neurons in the hidden layer is also determined after some experimentation: Figure 73 plots the *correct classification rates* $\times$ *number of neurons*. The graphic indicates that using 8 neurons yields the best rate.

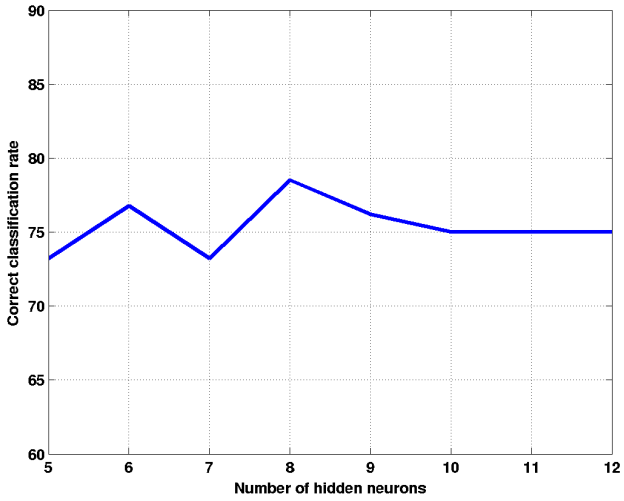


Figure 73 – ANN based on multilayer perceptron: setting the number of neurons in the hidden layer.

The second classifier considered was a Support Vector Machine (SVM), with a radial basis function (RBF) with the parameter width set to 0.01, after an analysis of the correct classification results. RBF has been chosen due to its capability of handling nonlinear relations between class labels and attributes, besides requiring reduced parameterization and resulting in high generalization [208]. The coefficient C for SVM is the soft margin parameter, which adjusts the pattern of the margin hyperplane according to the misclassified training instances. C was set by running several experiments with different values from 1 to 50 and analyzing the resulting correct classification rates, as depicted in Figure 74. It can be seen that the value $C = 35$ yields the best correct classification rate, hence it has been is chosen to adjust the SVM classifier.

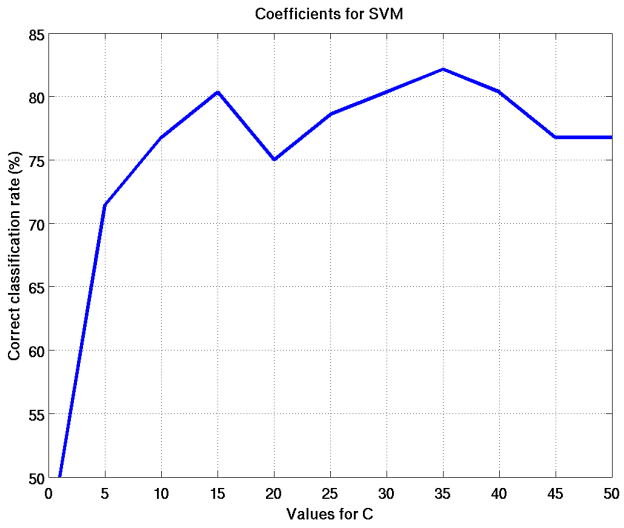


Figure 74 – SVM: setting the coefficient.

Finally, a K-Nearest Neighbor classifier was employed. Its use demands selecting a

distance function and the number of nearest neighbors K. The SID descriptor could consider any of the distance functions defined in Section 2.2.4, but the Euclidean distance has been picked for the experiments. The basic descriptor considers its associated dissimilarity measure defined in Section 2.2.4. Experiments have been conducted to set the value of K by performing classifications for various values of K and selecting the one that achieved the highest correct classification rate. Figure 75 presents a plot of the *number of neighbors* $\times$ *correct classification rates*, in which $K = 3$ is shown to produce the best correct classification rate.

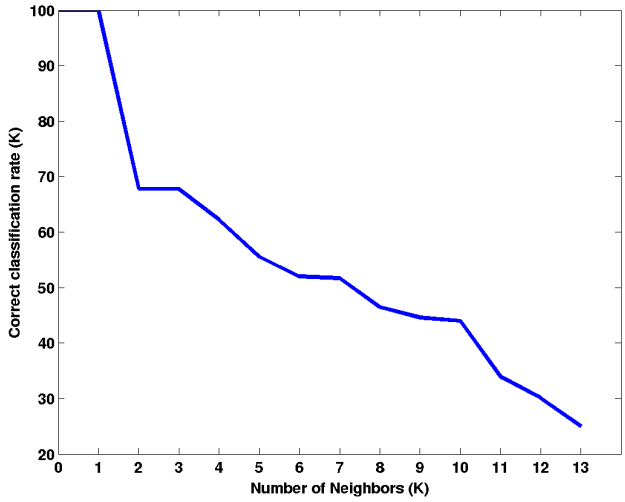


Figure 75 – K-NN: correct classification rates by varying the number of neighbors K.

The application of the basic green alga descriptor to the ANN-MLP, SVM, ID3 and the Customized Decision Tree (CDT) classifiers was restricted only to the basic shape features, since the CSS maxima are not quantitative measures that can be directly compared. Comparing CSS descriptors demands a matching process in which the resulting cost expresses their dissimilarity. Only the K-NN classifier is capable of incorporating the matching cost between two CSS descriptors.

Classifier performance

Classifiers have been validated with the leave-one-out strategy, in which each image in the dataset is used for testing and the remaining are used for training. This validation procedure ensures the generalization capability of trained classifiers.

Table 8 shows the correct classification rates obtained by the traditional classifiers K-NN, MLP, SVM and the decision trees ID3 and the proposed CDT, on the two alga image sets represented by both proposed descriptors, where each row expresses a particular combination. “N.A.” indicates that a classification task was not performed for the specific classifier and dataset configuration. Particularly, the CDT has not been tested on GA-123 due to practical limitations in creating the morphological characteristics table, as discussed in Chapter 6. This experiment reports that the classifiers SVM and MLP yielded the higher correct classifica-

tion rates on both shape sets. Such classification models have a peculiar learning process based on the optimization of decision boundaries, leading to more representative classification models. The ID3 decision tree achieved the lowest performance among the classifiers in set GA-56, while CDT performed better due to its customized formulation for identifying Selenastraceae algae. Generally, decision trees employ greedy algorithms in the learning process, so the obtained tree is not guaranteed to be optimal.

Table 8 – Correct classification results.

Image set	Descriptor	K-NN	MLP	SVM	ID3	CDT
GA-56	Basic	58.92%	73.21%	80.35%	**64.28%**	66.07%
GA-56	SID	**64.28%**	**78.57%**	**82.14%**	62.00%	**75.00%**
GA-123	Basic	64.22%	70.1%	72.87%	66.66%	N.A.
GA-123	SID	**69.91%**	**74.79%**	**78.07%**	**67.47%**	N.A.

Using the basic descriptor, K-NN was the only classifier to predict with higher accuracy in both green alga sets when compared to the description using SID. It is worth noting that the matching cost of CSS is included to the dissimilarity measure, so the description using the basic shape features and the CSS is more powerful. Moreover, in the set GA-123, the obtained accuracy does not present a significant difference because GA-123 has more images to compare with the nearest neighbors.

The analysis of the average accuracies in Table 8 allows to infer that the SID descriptor yielded a better discrimination capability between alga species by scoring more correctly classified instances. The exception occurs in the ID3 tree, in which the basic descriptor obtained a small advantage over SID. On set GA-123, accuracies were close for both descriptors. The K-NN classifier obtained the lowest correct classification rates, which suggests that learning approaches that employ optimization process are more suitable to the problem. However, it is worth noting that the dissimilarity measure might have a fundamental role since both descriptors employ the Euclidean distance for comparing instances.

Another interesting aspect shown in Table 8 is the differences between the correct classification rates over the two datasets. GA-56 is formed by ground-truth shapes and GA-123 is formed by shapes obtained from an automatic segmentation process. MLP held the performance, while SVM lost performance probably due to class imbalance in GA-123. The K-NN classifier improved the performance on GA-123 as more instances are available for the classification process.

Figure 76 depicts a comparison between the decision tree classifiers ID3 and the proposed CDT for each species. The x-axis refers to alga species (1,2,3,4,5,6 and 7) and the associated pair of bars show the correct classification rate (blue) and the F-Measure (red), with values in the y-axis. As this classification is multi-class, the F-Measure is more relevant, as a high number of true-negative hits may influence the analysis of classification results [209]. Figure 76(a) presents the classification of the GA-56 set described by SID (GA-56-SID). Poor performances are observed for species *Ankistrodesmus densus* (id 1), *Selenastrum bibraianum*

(id 2) and *Ankistrodesmus fusiformis* (id 7). Figure 76(b) illustrates the results of the CDT classifier on GA-56-SID, in which it can be seen that overall better F-Measure rates were obtained. Figure 76(c) shows the results of the CDT classifier on GA-56 using the Basic descriptor (GA-56-BAS). It can be seen that performance with SID was more uniform across alga species regarding F-Measure, whereas the Basic descriptor presented both high F-Measure rates, as in *Monoraphidium contortum* (id 6) and *Raphidocelis subcapitata* (id 3), and very low F-Measure rates, as reported for species *Selenastrum bibraianum* (id 2) and *Kirchineriella aperta* (id 4).

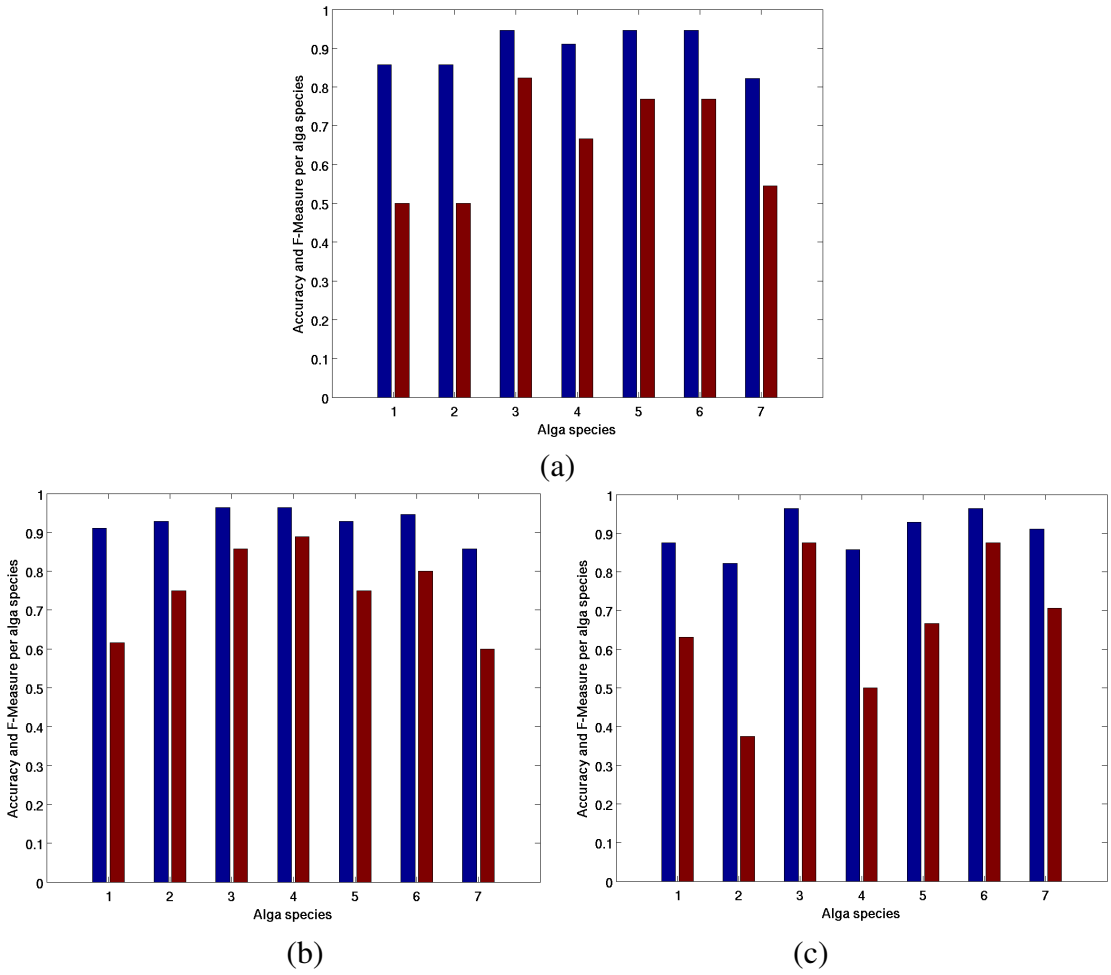
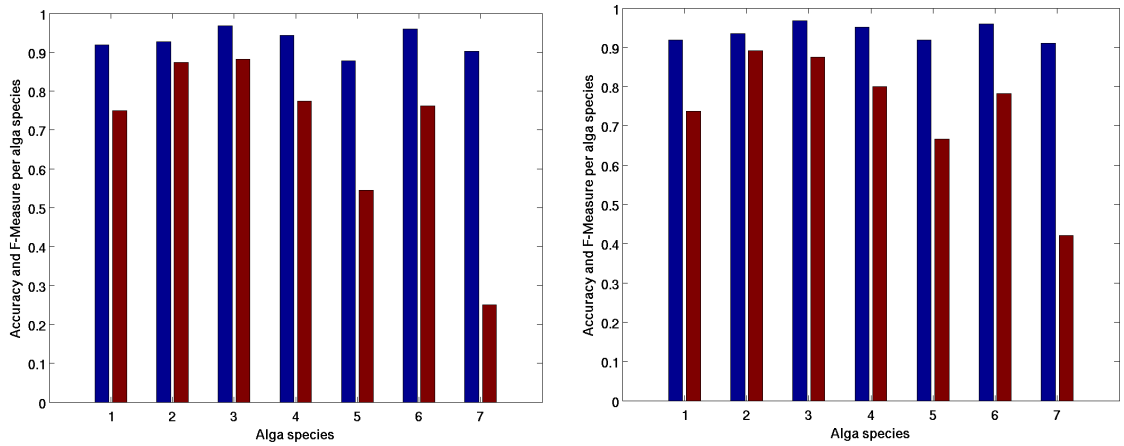


Figure 76 – Accuracy (blue bars) and F_1-Measure (red bars) for the set GA-56: (a) the ID3 decision tree on images described by the SID descriptor; (b) CDT on images described by the SID descriptor; (c) CDT on images described with the Basic descriptor.

Figure 77 describes and compares the classification results over each species obtained with the SVM and MLP classifiers, which yielded the best classification performances. In this case, the GA-123 set described by the SID descriptor (GA-123-SID) has been employed. Both classifiers performed poorly on *Ankistrodesmus fusiformis* (id 7) and *Monoraphidium griffithii* (id 5), mainly because these algae share similar morphological characteristics, such as in elongation and length. On the other hand, both classifiers achieved good differentiation of *Ankistrodesmus fusiformis* (id 7), *Selenastrum bibraianum* (id 2) and *Raphidocelis subcapitata* (id 3), which have visually dissimilar shape features. The main advantage of SVM over the MLP

classifier consists of a higher average precision and recall rates, but both achieved very good performances when classifying *Ankistrodesmus densus* (id 1) and *Raphidocelis subcapitata* (id 3).



(a) (b)

Figure 77 – Accuracy and F_1-Measure of the SVM and ANN-MLP classifiers on the seven alga species using the GA-123 set described by the Segments Intersection Descriptor: (a) ANN-MLP; (b) SVM.

Table 9 presents the confusion matrices for classifiers SVM and MLP for dataset GA-123-SID. Table 9 reveals the alga species that are similar to each other upon inspection of false negatives and false positive hits. Table 9(a) shows the confusion matrix for the SVM classifier, while Table 9(b) represents the confusion matrix from the classification results of the MLP. In both matrices, it can be seen that some images of *Monoraphidium griffithii* (id 5) have been classified as *Ankistrodesmus densus* (id 1) or as *Ankistrodesmus fusiformis* (id 7). The opposite situation also occurs, an evidence that these species share morphological characteristics: their elongated alga bodies vary in length according to the life stage and some deformations resulting from alga movement during image acquisition. Moreover, a few instances of *Raphidocelis subcapitata* (id 3) have been classified as *Kirchneriella aperta* (id 4) because both species have a C-shaped format, with a concavity that can vary in size or may be absent in colonies.

Table 9 – Confusion matrix for the SVM (a) and MLP (b) classifiers: rows indicate the real alga species and columns the predicted species.

GT/P	1	2	3	4	5	6	7
1	14	0	0	0	2	0	2
2	0	33	0	1	0	2	0
3	0	1	14	2	0	0	0
4	0	2	1	12	0	0	0
5	3	1	0	0	10	0	3
6	0	1	0	0	0	9	2
7	3	0	0	0	1	0	4

(a)

GT/P	1	2	3	4	5	6	7
1	15	0	0	0	2	0	1
2	1	31	0	3	0	1	0
3	0	1	15	1	0	0	0
4	0	2	1	12	0	0	0
5	4	0	0	0	9	0	4
6	0	1	1	0	1	8	1
7	2	0	0	0	4	0	2

(b)

In general, the experimental results can be summarized by the superior performance of SVM and MLP classifiers compared to K-NN and decision trees. Considering the characteristics

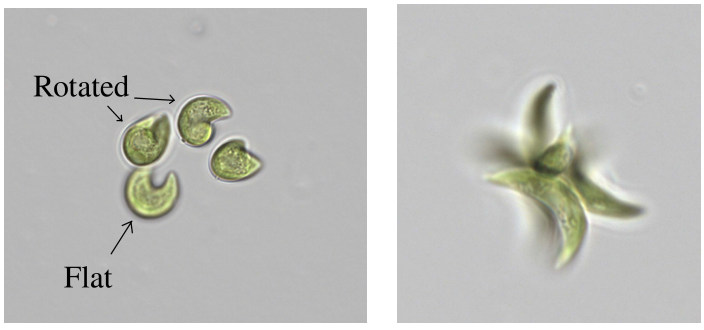
of green alga features and the nature of the taxonomical classification, approaches based on optimization processes have shown to be more appropriate for accurate and precise species recognition. However, the learning approaches of SVM and MLP present high computational cost. This is not a major concern as the green alga sets are typically small.

Furthermore, the experiments also indicated the better discrimination capability of the SID descriptor as compared with the descriptor composed by basic shape features. However, there are cases in which the highly similar morphological characteristics of alga species affect their description and consequently the correct classification hits. For instance, discriminating between *Ankistrodesmus fusiformis* and *Monoraphidium griffithii* is error prone because both have similar properties, such as elongation, thickness and small deformations, as illustrated in Figure 78. Figure 78(a) and 78(b) present straight long and thin green algae of *Monoraphidium griffithii* in their early cell stage. This means that a morphological characteristic that identifies an alga species may be also applicable to others, as shown in Figures 78(c-d), which represent samples of *Ankistrodesmus fusiformis*. According to the biologists, mistakes are common when attempting to identify such species.

(a) (b) (c) (d)

Figure 78 – Algae presenting highly similar shapes: (a-b) *Monoraphidium griffithii*; (c-d) *Ankistrodesmus fusiformis*.

Another factor that prevents better discrimination of alga species is their positioning and relative rotation in the microscope slide during the image acquisition, which captures a 3D real scene from the slide. Figure 79(a) depicts rotated and flat cells of *Kirchneriella aperta* in a zoomed in image and Figure 79(b) illustrates a *Selenastrum bribaianum* colony in which blurred regions indicate algae in a back position relative to the camera lens. This leads to incorrect detection of alga shapes of both species, which present a V-shaped geometry.



(a) (b)

Figure 79 – Rotated alga cells: (a) samples of *Kirchneriella aperta*; (b) samples of *Selenastrum bribaianum*.

Deriving a precise classification model is difficult due to the wide variety of shapes that can be found in samples from each alga species. Figure 80 depicts this situation in images of *Kirchneriella aperta*, in which Figures 80(a) and 80(b) refer to shapes of single cells and Figures 80(c) and 80(d) to colonies. These algae may also appear rotated in the 2D image, unlikely the standard shape profile considered in the taxonomy. Misclassified algae of this species are often predicted as belonging to *Raphidocelis subcapitata*, particularly the shape profiles 80(a) and 80(c).

(a) (b) (c) (d)

Figure 80 – The high variability of shapes of *Kirchneriella aperta*.

Increasing the number of images in the training set might improve species representativeness. However, it will also aggravate the problem of shape variance – as they move, several shapes can be observed from a single algae – and likely increase the occurrence of samples from different species sharing similar shape features. An ideal scenario would rely on acquiring microscope images in which alga cells present their standard and characteristic profile that is considered in the taxonomy.

Capturing these digital images is not straightforward since they must have uniform intensity patterns and a standard resolution for automatic processing and segmentation. Illumination conditions need to be similar to those of images employed in the experiments, for proper operation of the segmentation step. These physical properties of digital images directly affect the classification results, which justifies the several preprocessing steps prior to segmentation and the study of complex shape features for a more powerful description of alga morphological characteristics.

In the conventional procedure, an experienced biologist would need nearly one hour to perform the taxonomical identification of one alga organism from an image, as it is necessary to manually compute measures from the alga body in the microscope view to obtain the required information to apply the identification keys. Thus, an automated classification can support the biologists as an alternative source for identifying species. A biologist can compare the identification resulting from the identification keys with the automatic classification to increase their confidence in the process. Nonetheless, for the biologists to conduct the automatic classification is not simple, since they are not experienced in setting classifier parameters or even selecting appropriate classifiers and building training sets. Moreover, in many difficult cases their knowledge could be useful to improve classification performance. This motivated additional studies

on visual incremental classification processes, in which the user participates in the classification by interacting with visualizations and performing several classification iterations to improve results.

5.2 Visual classification

Visual classification is a user-assisted process that employs visual exploration to support classification tasks in scenarios where the relevant relations between data instances are assumed to be unknown. This process relies on visualizations capable of presenting users with a meaningful graphical metaphor capable of revealing relevant data patterns. Inserting the user into the classification loop requires some interactive tools support user feedback to the classifier based on previous knowledge or preliminary analysis. This scenario allows them 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.

Biologists engaged on a visual classification process can rely on their knowledge to select digital images with representative alga species for building training sets. For that purpose, two requirements are necessary. First, the set of green alga images should be presented in an intuitive manner. Visualization techniques based on phylogenetic trees provide a powerful graphical metaphor that can convey data similarities as a hierarchy to guide the selection of representative instances. Second, the shape features should describe the morphological characteristics of green alga reliably, so that the visualization expresses the similarity relations between alga images correctly. This is important, since the features should be representative and the biologists' knowledge can match the information presented in the visualization or classification results.

Paiva et al. [2] proposed a visual classification methodology applicable to image collections with user participation. Such methodology has been validated on collections of real-world images and demonstrated to be effective for obtaining efficient classification models. Applying this methodology to the classification of green microalga is interesting, because incremental classification can be useful to biologists with limited knowledge in setting classifier parameters and building training sets. Moreover, biologists can associate their previous knowledge about the taxonomy to interpret the similarity relations in the tree visualization.

Furthermore, *Paiva et al.* developed the *Visual Classification System* (VCS), an environment that implements the visual classification methodology and supports several functionalities, such as other alternative point-placement visualizations and modules for creating training sets, learning classification models, classifying datasets, and evaluating classification results by means of a vizualization-based tool. VCS also incorporates tools for interacting with the visualizations, such as viewing the images associated with tree nodes, zooming, splitting tree-based visualizations, and a tool for manually labeling instances directly in the visualization.

However, VCS is a system targeted at data mining experts who are familiar with param-

eter settings required by classifiers, which is not the case of biologists. VCS and its underlying methodology have been taken as a starting point of a system for incremental classification of Selenastraceae species from images. The essence of the visual classification methodology is maintained, but some modifications have been introduced into VCS to specialize its usage by the biologists.

5.2.1 Proposed visual classification process

The steps in the modified visual classification methodology are described next.

Visualizing green alga images

The first step refers to the NJ tree visualization of green alga images, in which the associated feature vectors are used as input to the Neighbor Joining algorithm that builds the hierarchical tree structure. The NJ-tree visualizations allow biologists to inspect relationships between images by zooming into branches and observing groups and their distribution in tree branches. They can use the interactive tools, for example to inspect the images associated with tree nodes, zoom in on selected branches and select individual nodes or branches.

Examples of possible analysis conducted on the tree visualizations are illustrated in Figures 81 and 82. Figure 81 shows a visualization of the GA-56-BAS data set. The tree has several branches, but class discrimination is relatively poor, as indicated by the fact that most branches have mixed species. A visual observation of the branches suggests that elongated algae *Ankistrodesmus densus* (blue nodes) and *Ankistrodesmus fusiformis* (dark red nodes) are well grouped, as well as the species characterized by small rounded cells *Kirchneriella aperta* (moss green) and *Raphidocelis subcapitata* (green). However, instances of *Selenastrum bibraianum* (cyan) and of *Monoraphidium contortum* (orange) appear at different branches of the tree, indicating a deficient description of their morphological characteristics.

Figure 82 presents the NJ-tree visualization of the GA-56-SID dataset. A global analysis of the visualization shows a clear segregation between species of algae that have rounded or elongated algae, that have been placed at branches to the left and to the right regions of the visualization, respectively. On the other hand, a local analysis of specific tree branches shows a high similarity between *Kirchneriella aperta* (moss green nodes) and *Raphidocelis subcapitata* (green nodes), *Monoraphidium contortum* (orange) and *Ankistrodesmus fusiformis* (dark red), and *Monoraphidium griffithii* and *Ankistrodesmus densus*. Unlike the visualization of Figure 81, instances of *Selenastrum bibraianum* (cyan) are placed in the middle regions of the tree.

An interesting aspect of the NJ-tree visualizations is the relation of the tree hierarchy and the classification results. Generally, a classifier works in a “black box” mode, in which the learning process is not visible to users, hindering the analysis of how such classification results were obtained. For instance, consider the confusion matrix in Table 10 relative to the MLP classifier applied to the same GA-56-SID data depicted in the visualization in Figure 82.

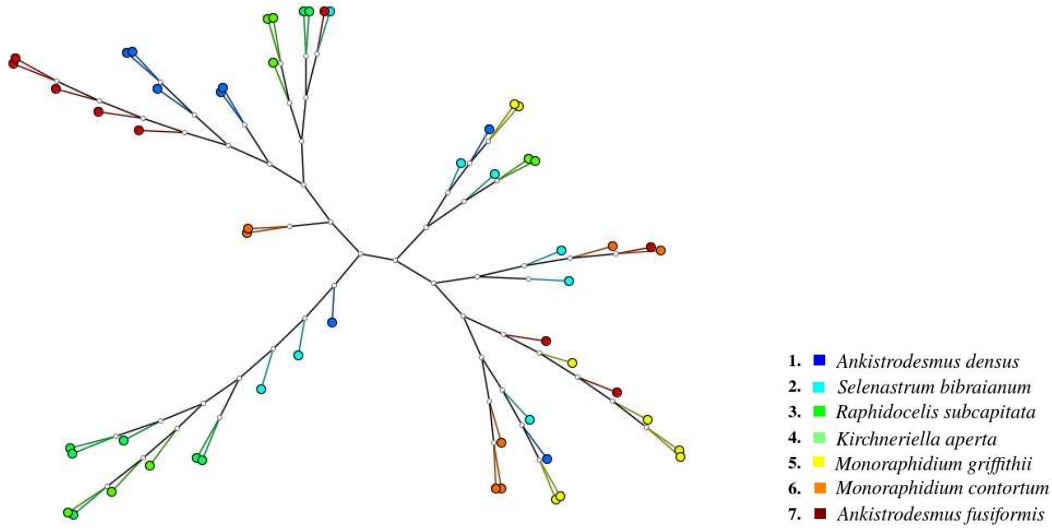


Figure 81 – NJ-tree visualization of the GA-56-BAS dataset (basic green alga descriptor).

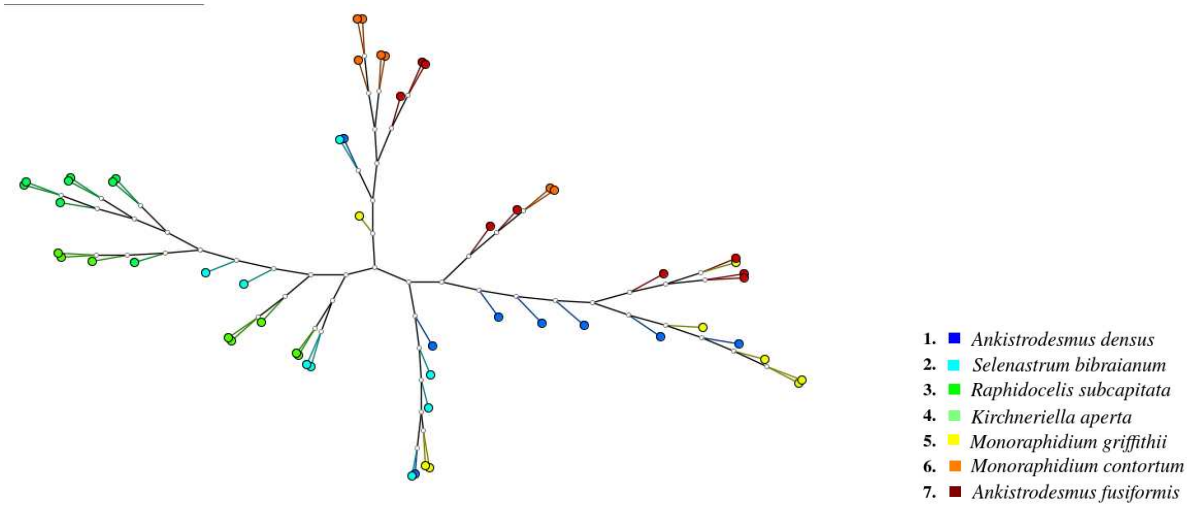


Figure 82 – NJ-tree visualization of the GA-56-SID dataset.

It can be seen that the most misclassified instances are from species *Selenastrum bibraianum* (cyan nodes), which had 2 instances incorrectly predicted as *Kirchneriella aperta* (moss green nodes). Inspecting the visualization, as depicted in Figure 83, it can be seen that both species appear either in the same branch or in close branches, indicating that their shape features are similar. Although some species, such as *Ankistrodesmus densus* (blue nodes) and *Monoraphidium griffithii* (yellow nodes) achieved high true-positive rates, the NJ-tree visualization does not segregate them well, suggesting that they present similar shape features (both are characterized by elongated and thick cells).

The NJ-tree and the classification algorithms use as input different dataset representations, i.e., NJ-tree considers the pairwise distances among data instances, while the classifiers (except K-NN) consider the feature vectors describing the instances. In some classification models, how instances are compared is not explicit to users, which explains why classifiers are as-

Table 10 – Confusion matrix for the MLP based classifier.

Ground-Truth / Predicted	1	2	3	4	5	6	7
1. *Ankistrodesmus densus*	7	1	0	0	0	0	0
2. *Selenastrum bibraianum*	1	4	0	2	0	0	1
3. *Raphidocelis subcapitata*	0	0	7	1	0	0	0
4. *Kirchneriella aperta*	1	0	1	6	0	0	0
5. *Monoraphidium grifithii*	1	0	0	0	6	0	1
6. *Monoraphidium contortum*	0	0	0	0	0	7	0
7. *Ankistrodesmus fusiformis*	0	1	0	0	0	1	7

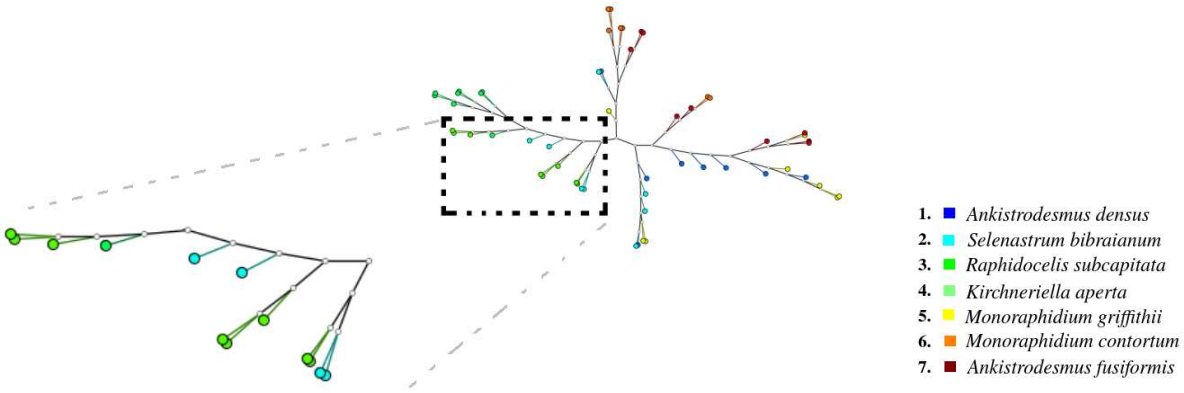


Figure 83 – Zooming into a branch with instances from species *Kirchneriella aperta* (moss green nodes) and *Selenastrum bibraianum* (cyan nodes).

signed as “black box” predictive models. The dissimilarity measure has a fundamental role in expressing the data relationships. Therefore, the major relation between the information displayed by the NJ-tree and the information input to classifiers is that they come from the same source, the data attributes, and a suitable dissimilarity measure can preserve such patterns. Thus, considering such scenario, tree-based visualizations can give users relevant information about data relations (in terms of dissimilarities) to support their assessment of the classifier results.

Visualizing and analyzing the data may happen at multiple moments during a visual classification process.

Building a training set

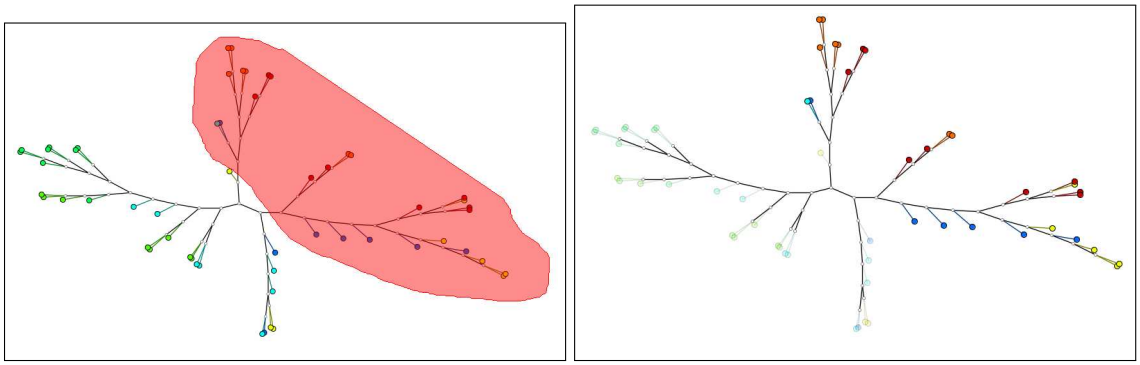
The classifiers applied to this problem are supervised approaches, which require training sets to learn the classification models. First, it is necessary to load a subset of images for which class information is previously known, and then generate the NJ-tree visualization. Using interactive tools for node selection, the biologist can analyze tree branches and their nodes and view the associated images.

During this process the biologist iteratively selects instances to compose the training set. Taking advantage of the tree structure, s/he can analyze and interpret the hierarchical relations between specific nodes and branches. the rationale is to select instances that are farther away from the center of the tree layout, because they are likely to be more representative of their

classes, according to the attributes (features) considered in creating the visualization. On the other hand, instances located near to the central area of the tree layout are not so representative of the features characteristic of their class.

There is no limit to the number of instances to be selected for the training set at each classification iteration. As the visual classification process is incremental, training sets can be updated as iterations succeed. If the classifier performance is not satisfactory, an updated training set with additional instances can be built for the next iteration. Otherwise, if the classification performance remains stable over several iterations even increasing the training set is increasing, the process can finish. Nonetheless, training sets should be representative, i.e., samples from all classes should be included for the learning process to derive a precise classification model.

Figure 84 shows the process of selecting instances for building a training set. The selection tool must be used, as indicated by the light red polygon in Figure 84(a). After selecting the instances, the visualization has its colors suppressed to highlight the selected instances, as shown in Figure 84(b).



(a) (b)

Figure 84 – Instance selection: (a) interactive tool for selecting nodes in the visualization; (b) selected instances are highlighted.

Learning the classifier and classifying green alga images

Once the training set is defined, a selected classifier can be employed to create a classification model, which may then be applied to the test dataset. The training and test instances must share the same feature space, i.e., instances in the test dataset must be described by the same descriptor used for the training instances. Figures 85(a) and 85(b) show NJ-tree visualization of the GA-56-SID image set: in (a) node color maps the actual species and in (b) node color maps the outcome of the classification.

An initial analysis of the correctly classified or misclassified instances can be done by means of a visual inspection of the NJ-tree depicting the classification results as compared with the ground truth. This may be a difficult task when handling larger data sets and many class labels.

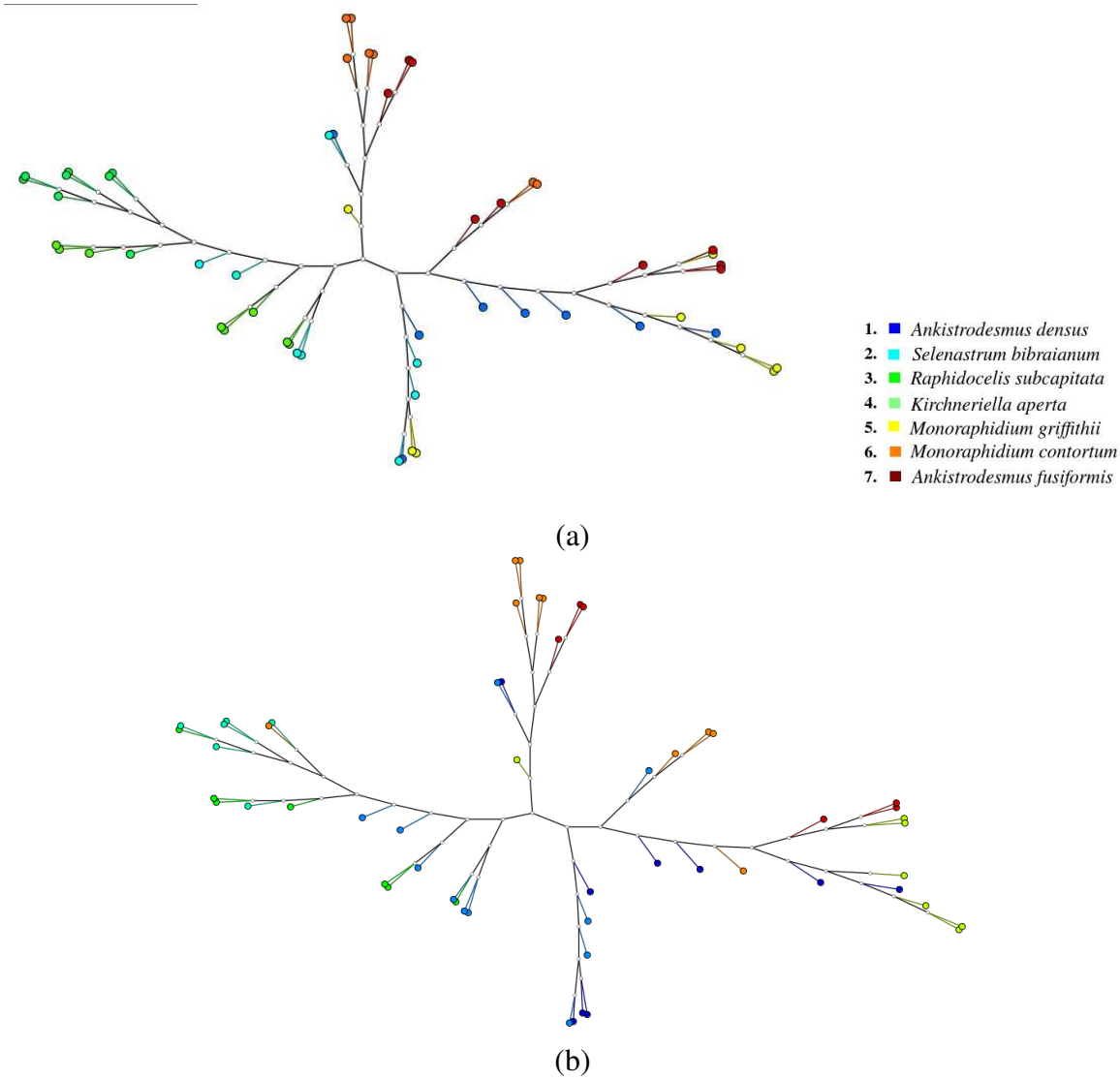


Figure 85 – Classification results of the GA-56-SID image set: (a) color maps ground-truth; (b) color maps classification results.

Evaluating the classification

Traditional metrics such as accuracy, precision, recall and confusion matrices can be used to analyze classification results. These metrics are informative, but they do not explicitly indicate the reasons that led to correct predictions or misclassified instances, a limitation that can be overcome using the NJ-tree visualization.

Originally implemented in VCS, the ClassMatching functionality takes advantage of the tree layout to support the interpretation of classification results: it displays the tree layout with nodes colored green or red, denoting respectively a correct or a misclassified instance. The user may focus at the misclassified instances to analyze how it relates with other instances and consider the possibility of updating the training set. Figure 86 illustrates ClassMatching for the GA-56-SID classification.

The classification process can be repeated when results are unsatisfactory or to improve

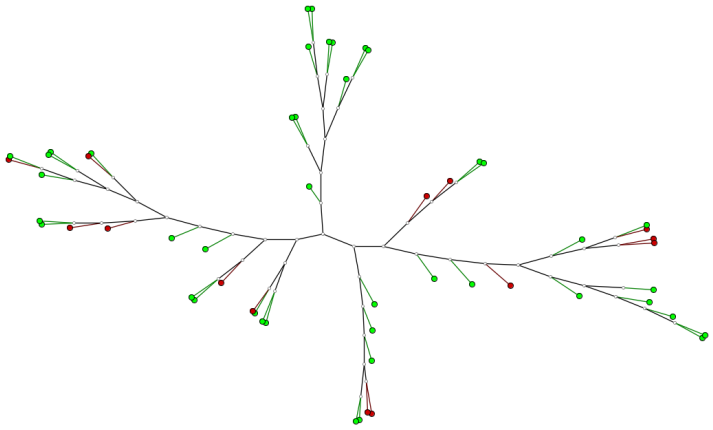


Figure 86 – Class Matching for assessing classification results: NJ-tree visualization in which red nodes indicate mismatches and green nodes are correct hits.

the performance of the classification model. Basically, other instances can be selected to for a new or an updated training set used to obtain a new classification model, and results can be assessed considering the metrics and the ClassMatching.

In the next section an experiment involving a biologist using the prototype is described to illustrate a real use case of incremental classification with the support of NJ-tree visualizations.

5.2.2 Experimental results

This section describes a case study depicting the biologist performing the visual classification process on set GA-123-SID. This experimentation employed a SVM classifier with the same settings reported in Section 5.1.2. The goal is to evaluate if the biologist would manage to create accurate classification models interpreting the NJ-tree visualizations by using his knowledge in favour to understand the similarity relations between green algae and to analyze classification results.

The biologist worked on a prototype version of VCS adapted to handle the specificities of this problem. Figure 87 depicts the interface of the prototype. The region labeled as “1” displays the NJ-tree visualizations and its associated interactions, e.g., checking individual nodes or selecting branches as indicated by the red polygon. Panel “2” presents the functionalities for activating the selection tool or the viewing module that displays the images associated to tree nodes. Moreover, the buttons to perform the classification are placed under the label “Tasks”, so the biologist can click on them and follow the instructions.

Panel “3” allows adjusting the tree-based visualizations, such as changing the nodes view to image thumbnails and modify the color mapping to show information about the morphological characteristics of the current dataset. Panel “4” shows a module for viewing the images associated to the selected instances in a separate frame. Some of the proposed visual resources existing in VCS were updated to facilitate use by the biologists.

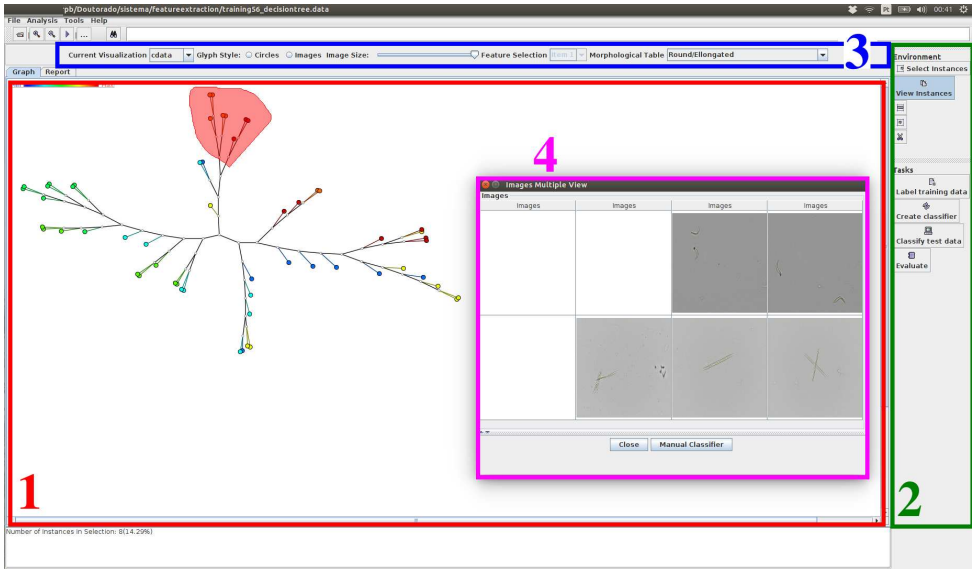


Figure 87 – The current layout of the adapted version of VCS for the biologists usage.

Before starting the case sudy, the biologist was presented a brief explanation on how to use the prototype and analyze the tree layout. The prototype's interface components and functions have been introduced, such as buttons, panels and menus. After a simple step-by-step demonstration of a visual classification, the biologist started and she managed to conduct the process by herself. At each rounded of the visual classification, the biologist faced some difficulties when attempting to select specific instances and the correct sequence for performing the steps of the incremental classification.

In the first iteration, the biologist selected 20 nodes at the top-right branches in the NJ-tree to compose the initial training set as depicted by the highlighted nodes in Figure 88(a). It is worth noting that this selection is not representative, as it does not include samples from all class labels. The biologist proceeded with the classification and the result is shown in Figure 88(b), in which the node colors of the NJ-tree depict the species predicted in the classification. Classification results are evaluated using the ClassMatching tool and the resulting layout is illustrated on Figure 88(c), which shows several misclassified instances towards the left side of the tree. Most of the correct hits refer to nodes from the training set, indicating that the classification model learned only the features of such instances. In this first iteration, 43 instances were correctly classified and 80 instances were misclassified, reaching an average precision of 24.32% and a recall of 34.96%. Such poor performance led the biologist to begin a new iteration to update the training set and repeat the learning process and classification.

The biologist started the second iteration by inspecting the NJ-tree visualization for interesting instances to be added to the existing current set. Her goal was to increase its coverage of classes, and she added 38 new samples from the leftmost branches. Figure 89(a) highlights the selected instances, and the resulting training set now includes samples from all species. The classification model has been updated and the classification results are shown in the NJ-tree

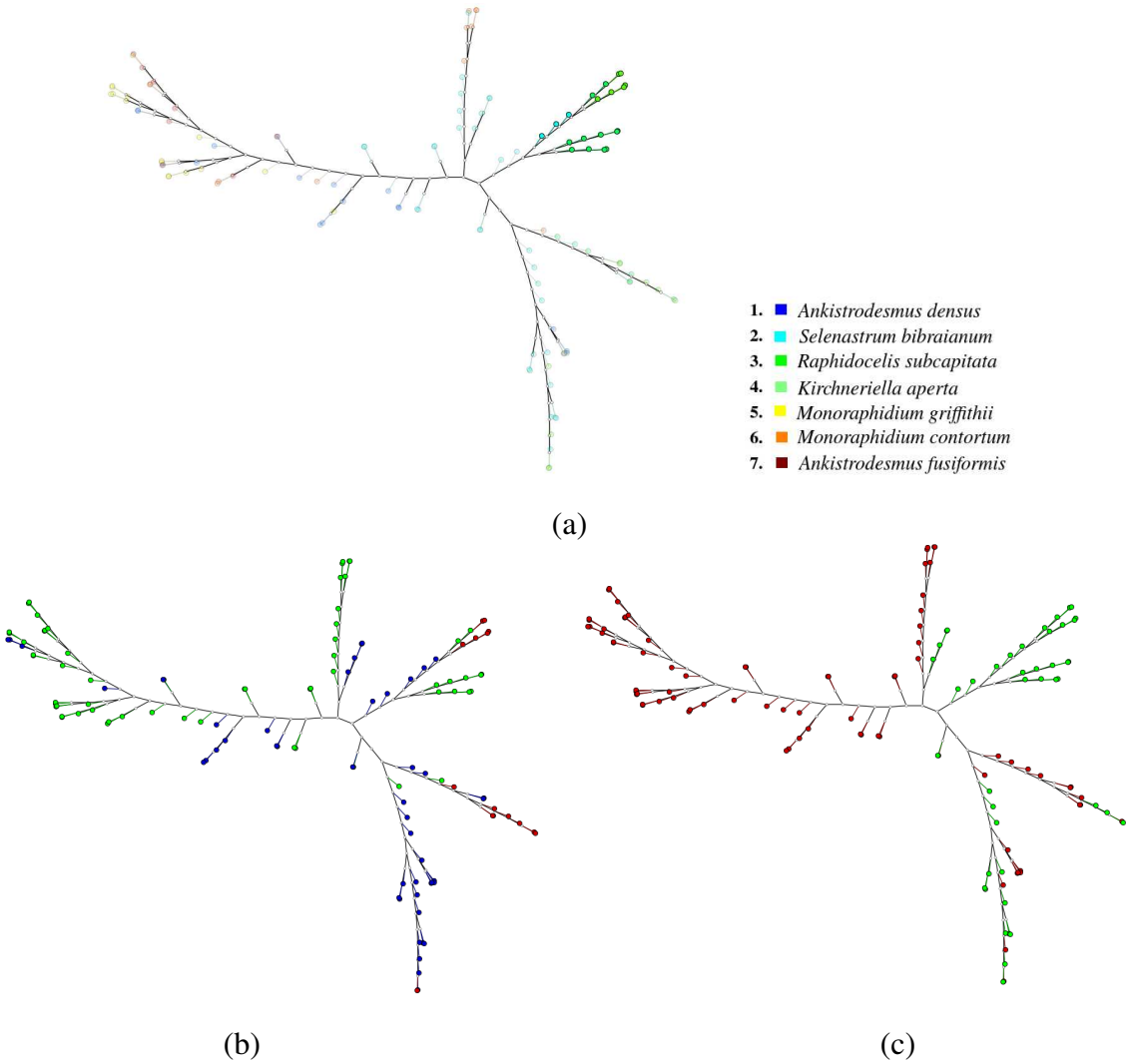


Figure 88 – First iteration of the visual classification: (a) 20 selected instances for the initial training set; (b) NJ-tree visualization representing the classification results; (c) ClassMatching view indicating 43 correct classifications.

depicted in Figure 89(b). The ClassMatching evidences 78 correct hits and 45 misclassified instances, as well as a precision of 64.78% and a recall of 36.64%. In this second rounded, classifier performance improved, since it employed a more appropriate training set. The ClassMatching (Figure 89(c)) tree reveals that most misclassified nodes are of samples from *Selenastrum bibraianum*, after analyzing their associated nodes in the NJ-tree representing the ground-truth. This led the biologist to perform a new iteration.

In this iteration, she selected 34 nodes in the NJ-tree visualization as depicted in Figure 90(a), so their associated instances are included into the current training set. The classification model is updated and the classification performed. Results are presented in Figure 90(b), and the ClassMatching view is shown in Figure 90(c), which indicates more instances correctly classified compared to the previous iteration: 92 correct hits and 31 misclassified instances. Thus, the training set obtained after 3 iterations is more representative, leading to the best classification performance.

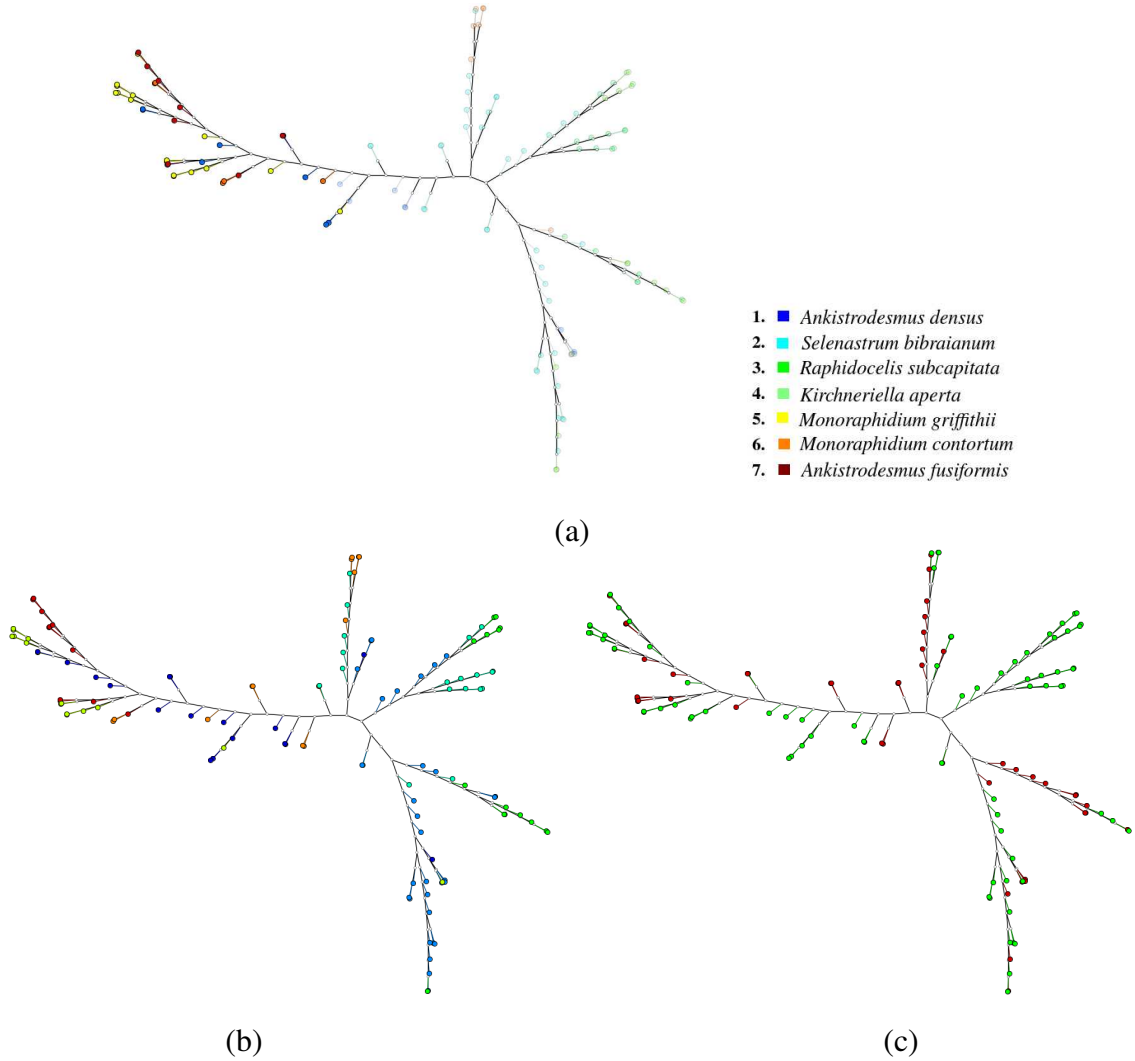


Figure 89 – Second iteration of the visual classification: (a) 38 selected instances for the training set; (b) NJ-tree visualization representing the classification results; (c) ClassMatching view indicating 78 correct classifications.

This case study suggests that incremental classification with the support of a visual exploration process is a potential tool to assist biologists in tasks of identification of green algae in images. The high variability of shapes makes the building of training sets particularly complex, as interesting instances may be missed and poor choices may lead to models with poor generalization capability. Thus, allowing biologists to use the visualizations to iteratively select instances for a suitable training set (repeating this procedure when the classification is not satisfactory) can increase the chances of successful classification.

Although the number of images available is not sufficient to justify a process for selecting instances to compose training sets, additional images can be gradually included. As the image sets become larger the selection of the most relevant instances turns out to be relevant due to computational issues. Furthermore, the biologist can specialize the classification task by considering subsets of green alga images or just some species by sampling instances in specific tree branches. For instance, s/he can choose to select just images of young cells for the training

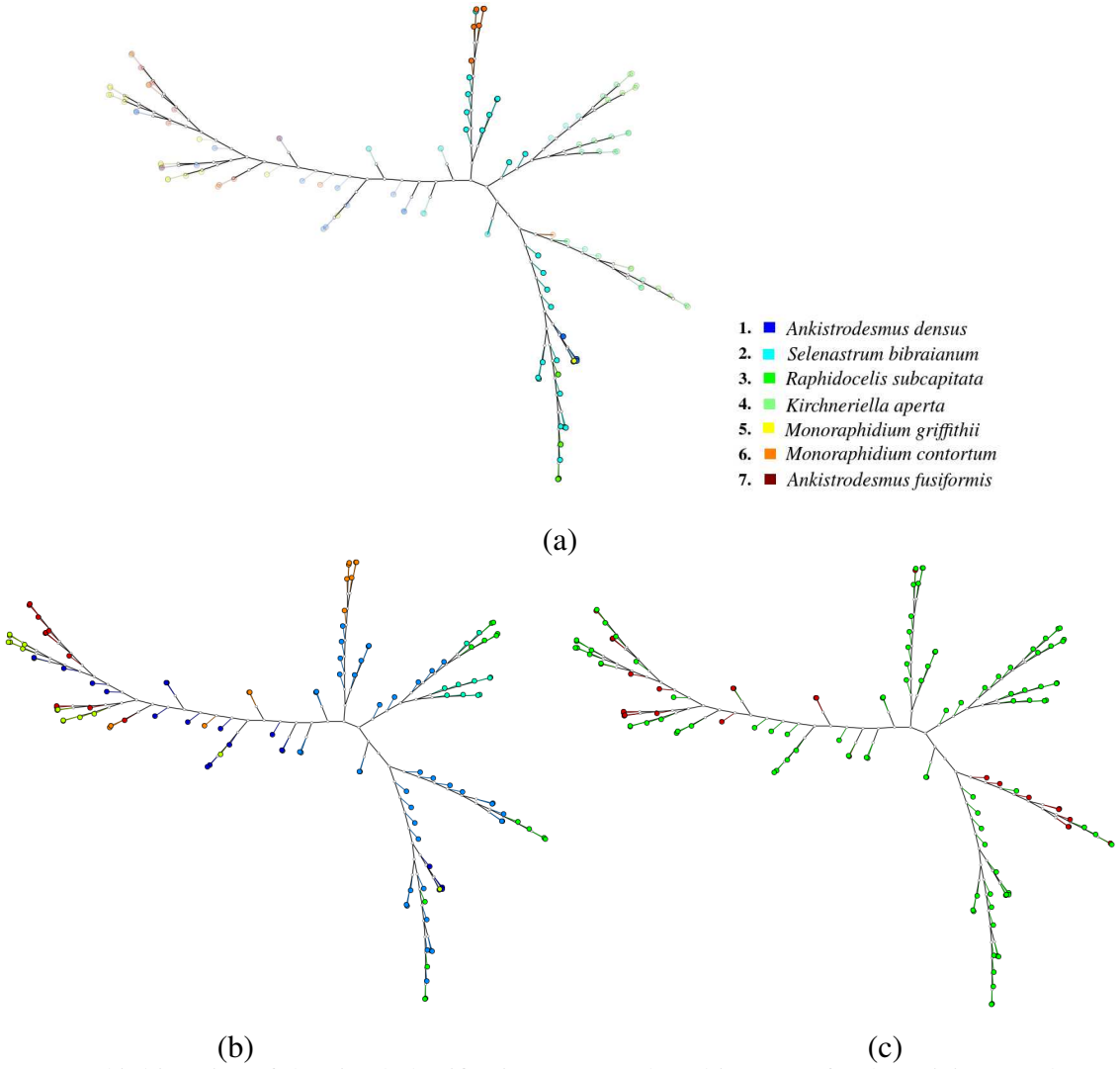


Figure 90 – Third iteration of the visual classification: (a) 34 selected instances for the training set; (b) NJ-tree visualization representing the classification results; (c) ClassMatching view indicating 92 correct classifications.

set, to train a classifier to predict species considering only their early life stages. Another example would be studying a specific genera, in which the training set could be built by selecting specific species of that genera that appear similar (close) in the visualization.

Another relevant concern is the usability of the prototype, in which the overall layout and interactive tools are fundamental to speed up the incremental classification process. After loading the data file containing the feature vectors, the biologist's first move was to check the functionality of each button and the effects of the interactive tools. Intuitive functionalities enable effective interactions with visualizations and other exploration analysis tools. Such aspect is relevant to prevent novice biologists using the prototype to face problems when attempting to understand the ordered steps to conduct the incremental and user-drive classification appropriately.

From the biologists point of view, using visualizations and incremental classification is interesting because it provides meaningful information that facilitates understanding the rela-

tionships between alga shape features. The tree layout groups images that are similar in regards to their morphological characteristics mapped to shape features. Analysis indicates that the tree layouts reflect the difficulties faced by taxonomists performing a conventional taxonomical identification of Selenastraceae in distinguishing between certain species. Finally, a complete system embedding the modules of image processing, shape-based feature extraction, visualization and incremental classification can support the taxonomical identification of Selenastraceae species by providing a "second opinion" to the traditional and manual procedures, with reduced effort. Thus, combining the conventional procedure with the proposed methodology in this work can reduce significantly the error rates in this task, and render the process more reliable and effective.

5.3 Final considerations

This chapter presented two methodologies for the taxonomical classification of Selenastraceae species on the proposed green alga descriptors. The first considers automatic strategies employing traditional classification algorithms, such as K-Nearest Neighbors, Support Vector Machines, Artificial Neural Network based on feed-forward Multilayer Perceptron and the ID3-based decision tree. Moreover, a customized decision tree has been devised with feedback from the biologists, in which its formulation and learning process resembles their conventional procedure towards classification.

The experimental results have shown superior performance of SVM and ANN classifiers when compared to the K-NN and the decision trees, since they employ an optimization process in the learning step. In general, those methods perform well for discriminating among alga genera, as well as classifying such organisms regarding their morphological characteristics. The classification of Selenastraceae has shown to be very difficult due to the similarities between alga shapes of different species and the large variability of shape formats. In the point of view of biologists, the results are satisfactory since the automatic classifiers make wrong prediction in the same way as taxonomists commit mistakes in the conventional identification procedure.

An alternative methodology relies on a visual classification process which allows users to gradually refine and improve the classification results. This strategy employs Neighbor Joining tree visualizations so that user knowledge can be employed to interpret and analyze the patterns and similarity relations between green alga images. Basically, the proposed visual incremental classification consists of the following steps: visualize the training set, interact with the visualization to select interesting instances to build the training set, learn the classification model, perform the classification and evaluate its performance using the ClassMatching tool. This iterative process proceeds until a representative and accurate classifier is obtained.

A biologist with solid knowledge on the Selenastraceae family performed the visual classification process on an adapted version of the Visual Classification System. Three itera-

tions were performed and she managed to increase the correct classification hits by updating the training set with representative instances. Finally, biologists provided a positive feedback, as they believe this approach can reduce their efforts while retaining accuracy in taxonomical identification tasks.

CHAPTER

6

CONCLUSIONS

This thesis presented a methodology for classifying green microalgae species of the Selenastraceae family from digital microscope images. The genomical studies on such family indicated that species which are morphologically similar may be molecularly very distinct, and also the opposite, with molecularly similar species showing diverse morphology. In addition, biologists recognize that current knowledge on the specific diversity and ecology of the Selenastraceae on a worldwide scale is limited. To clarify the many pending issues on genomical studies successfully, it is important to establish a common taxonomical basis using traditional approaches for species identification.

The conventional practices adopted by biologists is extremely time consuming, because it requires to obtain measures of alga morphological characteristics in the microscope and processing the identification keys for recognizing the species. Morever, the taxonomical identification is prone to error due to inconsistencies in Selenastraceae taxonomy, in which a morphological characteristic can occur in multiple species. Thus, it is important to improve the effectiveness of current practices adopted for analyzing morphological properties.

Therefore, this research afforded an excellent opportunity to apply image processing and pattern recognition techniques to a real problem in phycology. This thesis studied and proposed automated solutions to support the biologists in taxonomical identification processes of Selenastraceae alga species.

6.1 Contributions

The main contribution of this doctoral research is a computational methodology for species identification of algae from the Selenastraceae family in digital images. Most related work from the literature for automatic or semi-automated classification of algae or biological images are highly specialized to specific microorganisms. Thus, applying such methodologies to classify Selenastraceae species is likely to yield poor classification performance. As the au-

tomatic classification of Selenastraceae has not been explored using pattern recognition and image processing techniques, this research attempts to fill this gap. The proposed methodology follows the traditional pipeline of computer vision systems, typically comprising image preprocessing, segmentation, feature extraction and classification. Furthermore, a case study has also been conducted on a user-guided incremental classification with the support of visualizations.

Segmenting accurately green microalgae in digital images is fundamental to obtain representative shapes for a successful feature extraction. This means that a suitable segmentation method produces shapes that preserves the relevant structures suitable for the identification of each alga species. However, the segmentation of green alga images is particularly challenging, because images are naturally noisy, present heterogeneous colors and some artifacts that may be mistakenly recognized as alga cells. Some preprocessing techniques for noise suppressing and contrast enhancement have been employed to improve image quality. Generally, automated systems for identifying biological entities employ Gaussian filtering. However, to achieve image smoothing while preserving important boundary information of the target alga regions, the anisotropic diffusion filter was successfully applied [74].

Two approaches were developed for highly accurate segmentation of the images: a contour-based and a region-based. The former employs the level set method, in which a dynamic curve evolves through the image domain towards alga boundaries. The latter relies on a region growing principle in which seed pixels placed in candidate regions develop into a larger region, following a homogeneity criterion among neighboring pixels. Both methods employs probability distributions to represent the alga regions, computed from their intensity samples. Experiments showed that both methods were able to segment green algae accurately while preserving the relevant structures and shape properties. However, the region growing-based method presented more accurate segmentations. This is mainly due to the specialization of the region growing method for green alga images, as well as the employment of enhancement steps and the HSV color model. Although the method based on level set can be applied to a variety of images, obtaining the segmentation is extremely time consuming, which limits its application in practice. On the other hand, the methodology based on region growing proved to be efficient and effective for segmenting green alga images.

The feature extraction from green alga images is based on shape because the morphological characteristics are the main properties considered in the conventional taxonomical identification. The goal is to obtain a descriptor with high discrimination capability and low variance to rotation, scale and translations. The description of green algae is also challenging due to the wide diversity of shapes in each species, a consequence of various positions and life stages algae can appear in digital images. Initially, a few basic geometric shape features commonly employed for algae description in the literature have been applied for characterizing the algae. However, the experiments indicated that some descriptors are scale-dependent and cannot effectively capture the variations of shapes within and across species. Alternatively, the well-known Curvature Scale Space method was combined with such features, constituting a basic green

algae descriptor, which aims to improve the discrimination capability among species. Experiments showed that this descriptor was capable of distinguishing between some algae genera which are morphologically dissimilar. In addition, as CSS requires a matching process for measuring the dissimilarity among two feature vectors, it is complicated to incorporate the CSS maxima in the same way as other shape features into traditional classifiers, except K-NN.

A new and robust shape descriptor, called Segments Intersection Descriptor, has been proposed for capturing specific patterns of green alga shapes, such as isolated cell format (elongated or rounded) and colony geometry (rounded or stellate). The intuitive formulation consists of growing a circumference from the shape's centroid and detecting the overlapping pixels to the shape to create two signatures. Then, statistical measures are computed from these signatures to obtain a descriptor with low variance to rotation, scale and translation. Experiments have proved the high capability of SID in representing and describing alga shapes and other kinds of shapes.

The classification of green algae faces the same problems of the taxonomical identification since some Selenastraceae species share common morphological characteristics, hindering to derive a precise and robust classification model. In addition, the choice of the appropriate classifier is extremely complex due to the peculiar nature of green alga features. The criteria for selecting classifiers was based on the generalization capability and how intuitive the mathematical formulation of traditional classifiers best fits to the nature of green alga images. Classification methods, such as Artificial Neural Network (ANN), Support Vector Machine (SVM), K-Nearest Neighbors and a decision tree based on the ID3 have been applied to the green alga images for the taxonomical classification. Furthermore, a specialized decision tree that resembles the taxonomical identification of Selenastraceae have been proposed with the biologists' colaboration. Experiments have been performed using two green alga sets and for the proposed green alga descriptors. The results reported that the best classifiers, ANN and SVM, achieved accuracy rates around 77% and 71% for the SID-based descriptor, while the basic green algae descriptor presented a lower performance. The experiments indicated that some alga species share morphological characteristics or present similar shape features, such as between alga cells of *Ankistrodesmus fusiformis* and *Monoraphidium griffithii*.

Finally, an incremental classification strategy with the support of visualizations and user participation has been applied to the taxonomical identification of green algae. The main motivation is that biologists can not be experienced with automatic classifiers, since it requires to build training sets and to interpreting classification results. Using visualizations based on phylogenetic trees (the Neighbor Joining algorithm), an experiment was performed with a biologist conducting a visual classification in a set of green alga images. Using a prototype that is still under development, a biologist was able to interpret the classification results and to analyse the similarity relations among alga species in the visualization. Moreover, the biologist increased the accuracy of the classification over multiple iterations. The feedback on this strategy was positive, as well as the use of tree-based visualizations, as grouping of alga species were perceived

by taxonomists as similar to their analysis in the standard taxonomical identification process for the Selenastraceae.

6.2 Limitations

During the modeling and experimentation of the proposed techniques for image segmentation, shape-based feature extraction and classification, the following limitations were identified:

1. Although achieving high accuracy on segmentation, the proposed methodologies occasionally fail on low contrast digital images or when alga cell bodies are highly transparent. Although performed in the HSV model, segmentation remains challenging under these conditions.

2. The level set method can be applied to several kinds of images, but it is extremely time consuming since it requires an optimization process to evolve the dynamic curve towards alga boundaries. The region growing algorithm is faster than the level set method, but it is specialized to this particular application or to images with similar intensity patterns.

3. The Segment Intersection Descriptor does not appropriately characterize the local shape patterns or small subregions related to its area. Recognition may be impaired when shapes from different categories can only be distinguished by subtle details. Such limitation arises mainly because this descriptor represents the shape as a signature from which statistical measures are computed for a qualitative and quantitative characterization.

4. Characterization of green alga images by the proposed descriptors is relatively compromised due to the wide shape variations found within each species. The Selenastraceae taxonomy considers some shape formats by observing the morphological characteristics. However, in digital images, alga shapes might be rotated or overlapped since the organisms are constantly moving in the microscope slide during image acquisition.

5. The mapping from morphological characteristics to shape features, a strategy employed in both green alga descriptors, is relatively subjective since it requires knowledge about the shape's geometrical properties. In addition, it is required to analyze whether the mapped values (numerical attributes) can group similar instances regarding a specific morphological characteristic.

6. The proposed techniques for segmentation and feature extraction present some handcrafting because they were devised specifically for green alga images. As this research refers to the application of image processing, pattern recognition and visual data mining techniques to the taxonomical identification of green alga species, the proposed techniques have been specialized aiming to attain the highest correct classification rates.

7. The customized decision tree can only be applied to the green alga images described by the proposed descriptors. Although this decision tree has been manually designed to resemble the identification keys procedure, the inclusion of new species would require additional efforts to re-modeling the tree, i.e., re-arranging the nodes and the overall tree structure.

8. A more extensive evaluation of the visual incremental classification approach is required. Only one biologist was involved in the case study and the experiments only considered the scenario of selecting suitable instances to compose training sets for the adopted classifiers. This is due to the scarsity of experts available with solid knowledge on the Selenastraceae family and by the fact that selecting instances in phylogenetic trees explores better the advantages of visualizations, as it is more intuitive for the target users.

9. It is difficult to compare the conventional procedure conducted by biologists against the proposed classification approaches for identifying green alga species. Sampling algae from strains, viewing in the microscope, acquiring the digital image, measuring the morphological characteristics and then performing the identification keys per each alga cell requires approximately one hour's work from a biologist. Thus, comparing the identification accuracy of the conventional procedure for the green algae set constituted by 123 images is unfeasible for the biologist.

6.3 Future work

Future work contemplates new ideas that have emerged during the research and discussions, as well as to minimize and overcome the above mentioned limitations. The following proposals are suggested:

1. Deploy the prototype to the biologists, so they can actually perform multiple tasks of taxonomical classification of green alga images using the new tools provided. The prototype is still under development, with only the visual classification module available. The techniques for image processing and feature extraction were implemented in a separate environment and next steps comprise their inclusion into the prototype.

2. Explore the generalization capabilities of the proposed green alga descriptors to support new species that may be discovered as a result of biologists' research. This also includes additional studies on the formulation of the decision tree to allow the learning and prediction of additional alga species.

3. Investigate techniques for image enhancement so that the segmentation methods are able to process new green alga images with different illumination conditions when compared to the images handled in this research. Other kinds of images may be acquired from

different sources (microscopes) or can be captured by less experienced biologists with varying quality conditions, which can lead to poor classification performance.

4. Explore the relation between the NJ-tree visualization and the results obtained by classifiers in visual classification process. This means understanding when the layout quality is relevant or how it affects the classifier performance.

5. Study additional measures to improve the description of local shape patterns in order to improve discrimination capability. Moreover, it is also interesting to apply this descriptor to more challenging shape sets from other related knowledge domains, such as biomedicine and botany.

6. Evaluate possible approaches for integrating the solution developed for morphological characterization with complementary information provided by biologists, such as genomical information, for a more powerful classification of the Selenastraceae species.

6.4 Published papers

The following scientific papers have been published so far as a result of this research:

- Borges, V. R. P., Oliveira, M. C. F., Silva, T. G., Vieira, A. A., and Hamann, B., "Region growing for segmenting green microalgae images", to appear on IEEE/ACM Transactions on Computational Biology and Bioinformatics.

- Borges, V. R. P., Hamann, B., Silva, T. G., Vieira, A. A., and Oliveira, M. C. F., "A highly accurate level set approach for segmenting green microalgae images", IEEE Conference on Graphics, Patterns and Images (28^{th} SIBGRAPI), pp. 87-94, Salvador (Bahia), 2015.

- Borges, V. R. P., Hamann, B., Silva, T. G., Vieira, A. A., and Oliveira, M. C. F., "Feature Extraction and Interactive Visualization to Assist Green Algae Taxonomic Classification", Workshop of Works in Progress, IEEE Conference on Graphics, Patterns and Images (26^{th} SIBGRAPI), Arequipa (Peru), p. 1-4, 2013.

- Borges, V. R. P., Silva, T. G., Vieira, A. A. H and Oliveira, M. C. F. "Visual Classification of Freshwater Green Algae Images", 4^{th} Workshop on Interactive Data Visualization, p. 1-2, 2013. *(short paper)*

This work was also presented in the following conferences in phycology:

- Borges, V. R. P., Garcia, T. S., Hamann, B., Vieira, A. A. H., Oliveira, M. C. F., "An automatic methodology for morphological-based taxonomical classification of Selenastraceae green microalgae", XVI Congresso Brasileiro de Ficologia, Parnaiba- PI, 2016. *(poster)*

- Borges, V. R. P., Garcia, T. S., Hamann, B., Vieira, A. A. H., Oliveira, M. C. F., "Visual Exploration and Feature Extraction of Digital Images of Freshwater Green Microalgae to Assist Taxonomic Classification Tasks", IV Latin American Congress of Algae Biotechnology & Workshop of the National Network of Marine Algae Biotechnology, Florianópolis - SC, 2013. *(poster)*

- Borges, V. R. P., Silva, T. G., Vieira, A. A. H., Oliveira, M. C. F., Batista Neto, J. E. S. "Visualização exploratória e extração automática de características de imagens digitais de algas verdes de água doce.", XIV Congresso Brasileiro de Ficologia, João Pessoa - PB, 2012. *(poster)*

BIBLIOGRAPHY

[1] R. A. Fischer, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, vol. 7, pp. 179–188, 1936.

[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] D. Eler, F. Paulovich, M. Oliveira, and R. Minghim, "Coordinated and multiple views for visualizing text collections," *International Conference on Information Visualization*, pp. 246–251, 2008.

[4] F. V. Paulovich, L. G. Nonato, R. Minghim, and H. Levkowitz, "Least square projection: A fast high-precision multidimensional projection technique and its application to document mapping," *IEEE Transactions on Visualization and Computer Graphics*, vol. 14, pp. 564–575, 2008.

[5] 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.

[6] L. K. Leow, L.-L. Chew, V. C. Chong, and S. K. Dhillon, "Automated identification of copepods using digital image processing and artificial neural network," *BMC Bioinformatics*, vol. 16, no. Suppl 18, p. S4, 2015.

[7] M. A. Mosleh, H. Manssor, S. Malek, P. Milow, and A. Salleh, "A preliminary study on automated freshwater algae recognition and classification system," *BMC Bioinformatics*, vol. 13, no. Suppl 17, p. S25, 2012.

[8] P. Coltelli, L. Barsanti, V. Evangelista, A. M. Frassanito, and P. Gualtieri, "Water monitoring: automated and real time identification and classification of algae using digital microscopy," *Environmental Science: Processes & Impacts*, vol. 16, no. 11, pp. 2656–2665, 2014.

[9] P. Drews-Jr, R. G. Colares, P. Machado, M. de Faria, A. Detoni, and V. Tavano, "Microalgae classification using semi-supervised and active learning based on gaussian mixture models," *Journal of the Brazilian Computer Society*, vol. 19, no. 4, pp. 411–422, 2013.

[10] S. J. Fernstad, J. Johansson, S. Adams, J. Shaw, and D. Taylor, "Visual exploration of microbial populations," *IEEE Symposium on Biological Data Visualization (BioVis)*, vol. 1, pp. 127–134, 2011.

[11] J. Hasenauer, J. Heinrich, M. Doszczak, P. Scheurich, D. Weiskopf, and F. Allgöwer, "A visual analytics approach for models of heterogeneous cell populations," *Journal on Bioinformatics and Systems Biology*, no. 1, pp. 1–13, 2012.

[12] A. P. Francisco, C. Vaz, P. T. Monteiro, J. Melo-Cristino, M. Ramirez, and J. A. Carriço, "Phyloviz: phylogenetic inference and data visualization for sequence based typing methods," *BMC Bioinformatics*, vol. 13, no. 1, p. 1, 2012.

[13] V. R. P. Borges, B. Hamann, T. G. Silva, A. A. Vieira, and M. C. F. Oliveira, "A highly accurate level set approach for segmenting green microalgae images," *28th Conference on Graphics, Patterns and Images (SIBGRAPI)*, pp. 87–94, 2015.

[14] C. Bernard and F. Rassoulzadegan, "Bacteria or microflagellates as a major food source for marine ciliates: Possible implications for the microzooplankton," *Marine ecology progress series. Oldendorf*, vol. 64, no. 1, pp. 147–155, 1990.

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

[16] A. Demirbas, "Use of algae as biofuel sources," *Energy conversion and management*, vol. 51, no. 12, pp. 2738–2749, 2010.

[17] E. W. Becker, "Micro-algae as a source of protein," *Biotechnology advances*, vol. 25, no. 2, pp. 207–210, 2007.

[18] K. Skjånes, C. Rebours, and P. Lindblad, "Potential for green microalgae to produce hydrogen, pharmaceuticals and other high value products in a combined process," *Critical reviews in biotechnology*, vol. 33, no. 2, pp. 172–215, 2013.

[19] M. Shiho, M. Kawachi, K. Horioka, Y. Nishita, K. Ohashi, K. Kaya, and M. M. Watanabe, "Business evaluation of a green microalgae botryococcus braunii oil production system," *Procedia Environmental Sciences*, vol. 15, pp. 90–109, 2012.

[20] M. Soheili and K. Khosravi-Darani, "The potential health benefits of algae and micro algae in medicine: a review on spirulina platensis," *Current Nutrition & Food Science*, vol. 7, no. 4, pp. 279–285, 2011.

[21] M. W. Fawley, M. L. Dean, S. K. Dimmer, and K. P. Fawley, "Evaluation the morphospecies concept in the selenastraceae (chlorophyceae, chlorophyta)1," *Journal of phycology*, vol. 42, no. 1, pp. 142–154, 2006.

[22] L. Krienitz, I. Ustinova, T. Friedl, and V. A. R. Huss, "Traditional generic concepts versus 18s rrna gene phylogeny in the green algal family selenastraceae (chlorophyceae, chlorophyta)," *Journal of Phycology*, vol. 37, no. 5, pp. 852–865, 2001.

[23] P. D. Hebert, A. Cywinska, S. L. Ball *et al.*, "Biological identifications through dna barcodes," *Proceedings of the Royal Society of London B: Biological Sciences*, vol. 270, no. 1512, pp. 313–321, 2003.

[24] G. W. Saunders, "A dna barcode examination of the red algal family dumontiaceae in canadian waters reveals substantial cryptic species diversity. 1. the foliose dilsea-neodilsea complex and weeksia this paper is one of a selection of papers published in the special issue on systematics research." *Botany*, vol. 86, no. 7, pp. 773–789, 2008.

[25] J. Komárek, G. Huber-Pestalozzi, and B. Fott, "Das phytoplankton des süßwassers: Systematik und biologie. teil 7: Hälfte 1. chlorophyceae (grünalgen), ordnung: Chlorococcales," *IEEE Transactions on Visualization and Computer Graphics*, vol. 1, 1983.

[26] L. Krienitz and C. Bock, "Present state of the systematics of planktonic coccoid green algae of inland waters," *Hydrobiologia*, vol. 698, no. 1, pp. 295–326, 2012.

[27] R. Hoshina, "Dna analyses of a private collection of microbial green algae contribute to a better understanding of microbial diversity," *BMC Research Notes*, vol. 7, no. 1, p. 592, 2014.

[28] N. Santhi, C. Pradeepa, P. Subashini, and S. Kalaiselvi, "Automatic identification of algal community from microscopic images," *Bioinformatics and Biology Insights*, vol. 7, p. 327, 2013.

[29] S. Ohnuki, S. Nogami, S. Ota, K. Watanabe, S. Kawano, and Y. Ohya, "Image-based monitoring system for green algal haematococcus pluvialis (chlorophyceae) cells during culture," *Plant and Cell Physiology*, p. pct126, 2013.

[30] S. Wienert, D. Heim, K. Saeger, A. Stenzinger, M. Beil, P. Hufnagl, M. Dietel, C. Denkert, and F. Klauschen, "Detection and segmentation of cell nuclei in virtual microscopy images: a minimum-model approach," *Scientific reports*, vol. 2, 2012.

[31] L. Yuan, Y. F. Zheng, J. Zhu, L. Wang, and A. Brown, "Object tracking with particle filtering in fluorescence microscopy images: Application to the motion of neurofilaments in axons," *IEEE Transactions on Medical Imaging*, vol. 31, no. 1, pp. 117–130, 2012.

[32] A. Rizk, G. Paul, P. Incardona, M. Bugarski, M. Mansouri, A. Niemann, U. Ziegler, P. Berger, and I. F. Sbalzarini, "Segmentation and quantification of subcellular structures in fluorescence microscopy images using squassh," *Nature protocols*, vol. 9, no. 3, pp. 586–596, 2014.

[33] M. B. Meddens, B. Rieger, C. G. Figdor, A. Cambi, and K. van den Dries, "Automated podosome identification and characterization in fluorescence microscopy images," *Microscopy and microanalysis*, vol. 19, no. 01, pp. 180–189, 2013.

[34] M. C. F. Oliveira and H. Levkowitz, "From visual data exploration to visual data mining: A survey," *IEEE Transactions on Visualization and Computer Graphics*, vol. 9, no. 3, pp. 378–394, 2003.

[35] D. Keim, "Information visualization and visual data mining," *IEEE Transactions on Visualization and Computer Graphics*, vol. 8, no. 1, pp. 1–8, 2002.

[36] Y. Xu, C. Cheng, Y. Zhang, and D. Zhang, "Identification of algal blooms based on support vector machine classification in haizhou bay, east china sea," *Environmental Earth Sciences*, vol. 71, no. 1, pp. 475–482, 2014.

[37] D. Keim, "Visual exploration of large data sets," *Communications of the ACM*, vol. 44, no. 8, pp. 38–44, 2001.

[38] B. Waske, S. van der Linden, J. Benediktsson, A. Rabe, and P. Hostert, "Sensitivity of support vector machines to random feature selection in classification of hyperspectral data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 48, no. 7, pp. 2880 –2889, july 2010.

[39] 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.

[40] J. G. S. Paiva, W. R. Schwartz, H. Pedrini, and R. Minghim, "An approach to supporting incremental visual data classification," *IEEE Transactions on Visualization and Computer Graphics*, vol. 21, no. 1, pp. 4–17, 2015.

[41] L. A. Lewis and R. M. McCourt, "Green algae and the origin of land plants," *American Journal of Botany*, vol. 91, no. 10, pp. 1535–1556, 2004.

[42] S. Osher and J. A. Sethian, "Fronts propagating with curvature-dependent speed: algorithms based on hamilton-jacobi formulations," *Journal of Computational Physics*, vol. 79, no. 1, pp. 12–49, 1988.

[43] R. Adams and L. Bischof, "Seeded region growing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, no. 6, pp. 641–647, 1994.

[44] Y. Li, M. Horsman, N. Wu, C. Q. Lan, and N. Dubois-Calero, "Biofuels from microalgae," *Biotechnology progress*, vol. 24, no. 4, pp. 815–820, 2008.

[45] P. J. Keeling, G. Burger, D. G. Durnford, B. F. Lang, R. W. Lee, R. E. Pearlman, A. J. Roger, and M. W. Gray, "The tree of eukaryotes," *Trends in ecology & evolution*, vol. 20, no. 12, pp. 670–676, 2005.

[46] T. M. Mata, A. A. Martins, and N. S. Caetano, "Microalgae for biodiesel production and other applications: a review," *Renewable and sustainable energy reviews*, vol. 14, no. 1, pp. 217–232, 2010.

[47] R. J. Stevenson and Y. Pan, "Assessing environmental conditions in rivers and streams with diatoms," *The diatoms: applications for the environmental and earth sciences*, vol. 1, no. 1, p. 4, 1999.

[48] R. Andersen, "Diversity of eukaryotic algae," *Biodiversity & Conservation*, vol. 1, no. 4, pp. 267–292, 1992.

[49] W. Yee, "Microalgae from the selenastraceae as emerging candidates for biodiesel production: a mini review," *World Journal of Microbiology and Biotechnology*, vol. 32, no. 4, pp. 1–11, 2016.

[50] L. Krienitz and G. Klein, "Morphologie und ultrastruktur einiger arten der gattung monoraphidium (chlorellales) iii. monoraphidium terrestre bristol nov. comb." *Algological Studies/Archiv für Hydrobiologie, Supplement Volumes*, pp. 447–463, 1988.

[51] V. Eloranta, "The compound internal pyrenoid in cultured cells of the green algamonoraphidium griffithii (berkel.) komar.-legner." *Protoplasma*, vol. 99, no. 3, pp. 229–235, 1979.

[52] T. Miyashi, M. Kamata, and T. Mukai, "Oxygenation and [3+ 2]-cycloaddition of methylenecyclopropanes through electron donor-acceptor complexes with tetracyanoethylene by photoexcitation and in the dark," *Journal of the American Chemical Society*, vol. 108, no. 10, pp. 2755–2757, 1986.

[53] L. Krienitz, C. Bock, H. Nozaki, and M. Wolf, "Ssu rrna gene phylogeny of morphospecies affiliated to the bioassay alga "selenastrum capricornutum" recovered the polyphyletic origin of crescent-shaped chlorophyta," *Journal of Phycology*, vol. 47, no. 4, pp. 880–893, 2011.

[54] M. Hajibabaei, G. A. C. Singer, P. D. N. Hebert, and D. A. Hickey, "Dna barcoding: how it complements taxonomy, molecular phylogenetics and population genetics," *Trends in Genetic*, vol. 23, pp. 167–172, 2007.

[55] M. J. Colloff, "Taxonomy and identification of dust mites," *Allergy*, vol. 53, pp. 7–12, 1998.

[56] G. Huber-Pestalozzi, *Das Phytoplankton des susswassers: Systematik und Biologie*, ser. Das Phytoplankton des susswassers: Systematik und Biologie. Schweizerbart, 1938, no. v. 16, pt. 2, no. 1.

[57] G. Novarino, "A companion to the identification of cryptomonad flagellates (cryptophyceae= cryptomonadea)," *Phytoplankton and Equilibrium Concept: The Ecology of Steady-State Assemblages*, vol. 1, pp. 225–270, 2003.

[58] J. Davis, C. Martin, and D. Mackenzie, "A simple expert system for identifying plants in a local area," *Environmental Software*, vol. 5, no. 3, pp. 149 – 157, 1990.

[59] A. G. Comas, "Las chlorococcales dulciacuícolas de cuba.[the freshwater chlorococcids of cuba.] stuttgart: J. cramer. 192p., il," *Bibliotheca Phycologica*, 1996.

[60] F. Hindák, "New taxa and reclassifications in the chlorococcales (chlorophyceae)," *Preslia*, 1978.

[61] B. Fott, *Studies in phycology*. Schweizerbart, 1969.

[62] R. Rojíčková-Padrtová, B. Maršálek, and I. Holoubek, "Evaluation of alternative and standard toxicity assays for screening of environmental samples: selection of an optimal test battery," *Chemosphere*, vol. 37, no. 3, pp. 495–507, 1998.

[63] J. Ma, S. Wang, P. Wang, L. Ma, X. Chen, and R. Xu, "Toxicity assessment of 40 herbicides to the green alga raphidocelis subcapitata," *Ecotoxicology and Environmental Safety*, vol. 63, no. 3, pp. 456–462, 2006.

[64] R. C. Gonzalez, *Digital image processing*. Pearson Education India, 2008.

[65] E. Dougherty and J. Astola, *Nonlinear Filters for Image Processing*, ser. SPIE/IEEE Series on Imaging Science & Engineering. Wiley, 1999.

[66] A. Buades, B. Coll, and J. Morel, "A review of image denoising algorithms, with a new one," *Multiscale Modeling & Simulation*, vol. 4, no. 2, pp. 490–530, 2005.

[67] P. Perona and J. Malik, "Scale-space and edge detection using anisotropic diffusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 7, pp. 629 –639, jul 1990.

[68] J. Weickert, *Anisotropic Diffusion in Image Processing*, ser. ECMI Series. Teubner-Verlag, Stuttgart, 1996.

[69] M. Hajiaboli, "An anisotropic fourth-order diffusion filter for image noise removal," *International Journal of Computer Vision*, vol. 92, pp. 177–191, 2011.

[70] J.-S. Lee, "Digital image enhancement and noise filtering by use of local statistics," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-2, no. 2, pp. 165 –168, march 1980.

[71] M. Nitzberg and T. Shiota, "Nonlinear image filtering with edge and corner enhancement," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, no. 8, pp. 826 –833, aug 1992.

[72] M. Sonka, V. Hlavac, and R. Boyle, *Image processing, analysis, and machine vision*. Cengage Learning, 2014.

[73] C.-T. Lu and T.-C. Chou, "Denoising of salt-and-pepper noise corrupted image using modified directional-weighted-median filter," *Pattern Recognition Letters*, vol. 33, no. 10, pp. 1287–1295, 2012.

[74] C. Z. Barcelos and V. Pires, "An intelligent method for edge detection based on nonlinear diffusion," *Artificial Intelligence in Theory and Practice*, vol. 215, no. 1, pp. 329–338, 2008.

[75] V. Prasath, R. Pelapur, O. Glinskii, V. Glinsky, V. Huxley, and K. Palaniappan, "Multiscale tensor anisotropic filtering of fluorescence microscopy for denoising microvasculature," *12th International Symposium on Biomedical Imaging (ISBI)*, vol. 1, pp. 540–543, 2015.

[76] C. A. Z. Barcelos and V. Pires, "An automatic based nonlinear diffusion equations scheme for skin lesion segmentation," *Applied Mathematics and Computation*, vol. 215, no. 1, pp. 251–261, 2009.

[77] S. Gupta and S. G. Mazumdar, "Sobel edge detection algorithm," *International journal of computer science and management Research*, vol. 2, no. 2, pp. 1578–1583, 2013.

[78] S. S. Sannakki, V. S. Rajpurohit, V. Nargund, and P. Kulkarni, "Diagnosis and classification of grape leaf diseases using neural networks," *4th International Conference on Computing, Communications and Networking Technologies*, vol. 1, pp. 1–5, 2013.

[79] G. Aubert and P. Kornprobst, "Mathematical problems in image processing: partial differential equations and the calculus of variations," *Springer Science & Business Media*, vol. 147, 2006.

[80] W. F. Ames, *Numerical methods for partial differential equations*. Academic Press, 2014.

[81] S. Suganthi and S. Ramakrishnan, "Anisotropic diffusion filter based edge enhancement for segmentation of breast thermogram using level sets," *Biomedical Signal Processing and Control*, vol. 10, pp. 128–136, 2014.

[82] J. Canny, "A computational approach to edge detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. 6, pp. 679–698, 1986.

[83] J. Illingworth and J. Kittler, "A survey of the hough transform," *Computer Vision, Graphics, and Image Processing*, vol. 44, no. 1, pp. 87 – 116, 1988.

[84] D. Ballard, "Generalizing the hough transform to detect arbitrary shapes," *Pattern Recognition*, vol. 13, no. 2, pp. 111 – 122, 1981.

[85] D. Marr and E. Hildreth, "Theory of edge detection," *Royal Society of London Proceedings Series B*, vol. 207, pp. 187–217, 1980.

[86] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *International Journal of Computer Vision*, vol. 1, no. 4, pp. 321–331, 1988.

[87] V. Borges, M. Batista, and C. Barcelos, "A soft unsupervised two-phase image segmentation model based on global probability density functions," *IEEE International Conference on Systems, Man, and Cybernetics*, pp. 1687 –1692, 2011.

[88] J. Tang, "A color image segmentation algorithm based on region growing," *2nd International Conference on Computer Engineering and Technology*, vol. 6, pp. V6–634 –V6–637, april 2010.

[89] M. Spann, "A quad-tree approach to image segmentation which combines statistical and spatial information," *Pattern Recognition*, vol. 18, no. 3-4, pp. 257 – 269, 1985.

[90] M. A. G. Carvalho, A. C. B. Ferreira, and A. L. Costa, "Image segmentation using quadtree-based similarity graph and normalized cut," *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, vol. 6419, pp. 329–337, 2010.

[91] S. Beucher and C. D. M. Mathmatique, "The watershed transformation applied to image segmentation," *Scanning Microscopy International*, pp. 299–314, 1991.

[92] A. Bleau and L. Leon, "Watershed-based segmentation and region merging," *Computer Vision and Image Understanding*, vol. 77, no. 3, pp. 317 – 370, 2000.

[93] G. Hamarneh and X. Li, "Watershed segmentation using prior shape and appearance knowledge," *Image and Vision Computing*, vol. 27, no. 1-2, pp. 59 – 68, 2009.

[94] H. M. Sosik and R. J. Olson, "Automated taxonomic classification of phytoplankton sampled with imaging-in-flow cytometry," *Limnology and Oceanography: Methods*, vol. 5, no. 6, pp. 204–216, 2007.

[95] I. Dimitrovski, D. Kocev, S. Loskovska, and S. Džeroski, "Hierarchical classification of diatom images using ensembles of predictive clustering trees," *Ecological Informatics*, vol. 7, no. 1, pp. 19–29, 2012.

[96] H. Zheng, H. Zhao, X. Sun, H. Gao, and G. Ji, "Automatic setae segmentation from chaetoceros microscopic images," *Microscopy research and technique*, vol. 77, no. 9, pp. 684–690, 2014.

[97] N. Otsu, "A threshold selection method from gray-level histograms," *Automatica*, vol. 11, no. 285-296, pp. 23–27, 1975.

[98] P. N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*. Pearson, Addison Wesley, 2006.

[99] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*, 2nd ed. Wiley, 2001.

[100] R. O. Stehling, M. A. Nascimento, and A. X. Falcão, "A compact and efficient image retrieval approach based on border/interior pixel classification," *Proceedings of the eleventh international conference on Information and knowledge management*, vol. 1, pp. 102–109, 2002.

[101] G. Pass, R. Zabih, and J. Miller, "Comparing images using color coherence vectors," in *Proceedings of the fourth ACM international conference on Multimedia*. ACM, 1997, pp. 65–73.

[102] K. Roy and J. Mukherjee, "Image similarity measure using color histogram, color coherence vector, and sobel method," *International Journal of Science and Research (IJSR)*, vol. 2, no. 1, 2013.

[103] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Transactions on Systems, Man and Cybernetics*, vol. SMC-3, no. 6, pp. 610 –621, nov. 1973.

[104] V. S. Thakare and N. N. Patil, "Classification of texture using gray level co-occurrence matrix and self-organizing map," *International Conference on Electronic Systems, Signal Processing and Computing Technologies*, vol. 16, pp. 350–355, 2014.

[105] X. Qi, R. Xiao, C.-G. Li, Y. Qiao, J. Guo, and X. Tang, "Pairwise rotation invariant co-occurrence local binary pattern," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 11, pp. 2199–2213, 2014.

[106] M. M. Galloway, "Texture analysis using gray level run lengths," *Computer Graphics and Image Processing*, vol. 4, no. 2, pp. 172 – 179, 1975.

[107] A. B. Tosun and C. Gunduz-Demir, "Graph run-length matrices for histopathological image segmentation," *IEEE Transactions on Medical Imaging*, vol. 30, no. 3, pp. 721–732, 2011.

[108] A. R. Backes and O. M. Bruno, "A graph-based approach for shape skeleton analysis," *Image Analysis and Processing*, pp. 731–738, 2009.

[109] J. J. d. M. S. Junior and A. R. Backes, "Shape classification using line segment statistics," *Information Sciences*, vol. 305, pp. 349–356, 2015.

[110] M. Singh and D. D. Hoffman, "Natural selection and shape perception," *Shape Perception in Human and Computer Vision*, pp. 171–185, 2013.

[111] J. B. Florindo, A. R. Backes, M. de Castro, and O. M. Bruno, "A comparative study on multiscale fractal dimension descriptors," *Pattern Recognition Letters*, vol. 33, no. 6, pp. 798–806, 2012.

[112] A. C. Jalba, M. H. Wilkinson, and J. B. Roerdink, "Shape representation and recognition through morphological curvature scale spaces," *IEEE Transactions on Image Processing*, vol. 15, no. 2, pp. 331–341, 2006.

[113] M. Lauffer, F. Genty, J.-L. Collette, and S. Margueron, "Phytoplankton identification by combined methods of morphological processing and fluorescence imaging," *IEEE Workshop on Environmental, Energy and Structural Monitoring Systems (EESMS)*, vol. 1, pp. 131–135, 2015.

[114] R. S. Torres and A. X. Falcao, "Contour salience descriptors for effective image retrieval and analysis," *Image and Vision Computing*, vol. 25, no. 1, pp. 3–13, 2007.

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

[116] R. Jain, S. Murthy, P.-J. Chen, and S. Chatterjee, "Similarity measures for image databases," *IEEE International Conference on Fuzzy Systems*, vol. 3, pp. 1247 –1254, 1995.

[117] O. Kutz, F. Wolter, H. Sturm, N.-Y. Suzuki, and M. Zakharyaschev, "Logics of metric spaces," *ACM Transactions on Computational Logic*, vol. 4, no. 2, pp. 260–294, 2003.

[118] D. Zhang and G. Lu, "Evaluation of similarity measurement for image retrieval," *International Conference on Neural Networks and Signal Processing*, vol. 2, pp. 928 –931, dec. 2003.

[119] L. Song, D. Li, X. Zeng, Y. Wu, L. Guo, and Q. Zou, "ndna-prot: identification of dna-binding proteins based on unbalanced classification," *BMC Bioinformatics*, vol. 15, no. 1, p. 1, 2014.

[120] K. Dannemiller, K. Ahmadi, and E. Salari, "A new method for the segmentation of algae images using retinex and support vector machine," *IEEE International Conference on Electro/Information Technology*, vol. 1, pp. 361–364, 2015.

[121] A. Verikas, A. Gelzinis, M. Bacauskiene, I. Olenina, and E. Vaiciukynas, "An integrated approach to analysis of phytoplankton images," *IEEE Journal of Oceanic Engineering*, vol. 40, no. 2, pp. 315–326, 2015.

[122] A. Khataee, G. Dehghan, A. Ebadi, M. Zarei, and M. Pourhassan, "Biological treatment of a dye solution by macroalgae chara sp.: Effect of operational parameters, intermediates identification and artificial neural network modeling," *Bioresource technology*, vol. 101, no. 7, pp. 2252–2258, 2010.

[123] K. Schulze, U. M. Tillich, T. Dandekar, and M. Frohme, "Planktovision–an automated analysis system for the identification of phytoplankton," *BMC Bioinformatics*, vol. 14, no. 1, p. 115, 2013.

[124] F. García-Camacho, L. López-Rosales, A. Sánchez-Mirón, E. Belarbi, Y. Chisti, and E. Molina-Grima, "Artificial neural network modeling for predicting the growth of the microalga karlodinium veneficum," *Algal Research*, vol. 14, pp. 58–64, 2016.

[125] J. J. Hopfield, "Artificial neural networks," *IEEE Circuits and Devices Magazine*, vol. 4, no. 5, pp. 3–10, 1988.

[126] D. W. Patterson, *Artificial neural networks: theory and applications*. Prentice Hall PTR, 1998.

[127] D. Graupe, *Principles of artificial neural networks*. World Scientific, 2013, vol. 7.

[128] R. L. De Mántaras, "A distance-based attribute selection measure for decision tree induction," *Machine learning*, vol. 6, no. 1, pp. 81–92, 1991.

[129] N. Bhargava, G. Sharma, R. Bhargava, and M. Mathuria, "Decision tree analysis on j48 algorithm for data mining," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 3, no. 6, 2013.

[130] S. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 21, no. 3, pp. 660–674, 1991.

[131] R. C. Barros, M. P. Basgalupp, A. C. De Carvalho, and A. A. Freitas, "A survey of evolutionary algorithms for decision-tree induction," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 3, pp. 291–312, 2012.

[132] J. R. Quinlan, "Induction of decision trees," *Machine learning*, vol. 1, no. 1, pp. 81–106, 1986.

[133] K. K. R. Nynalasetti, G. Varma, and M. N. Rao, "Classification rules using decision tree for dengue disease," *International Journal of Research in Computer and Communication Technology*, vol. 3, no. 3, pp. 340–343, 2014.

[134] A. T. Azar and S. M. El-Metwally, "Decision tree classifiers for automated medical diagnosis," *Neural Computing and Applications*, vol. 23, no. 7-8, pp. 2387–2403, 2013.

[135] B. Lakshmipathi and G. Kousalya, "Identifying genetic defected dna by using c4. 5 based binary decision tree classifier," *Middle-East Journal of Scientific Research*, vol. 23, no. 6, pp. 1197–1203, 2015.

[136] J. Davis and M. Goadrich, "The relationship between precision-recall and roc curves," *23rd International Conference on Machine Learning*, pp. 233–240, 2006.

[137] S. Card, J. Mackinlay, and B. Shneiderman, *Readings in Information Visualization: Using Vision to Think*, ser. The Morgan Kaufmann Series in Interactive Technologies. Morgan Kaufmann, 1999.

[138] C. Ware, *Information Visualization: Perception for Design*, ser. The Morgan Kaufmann Series in Interactive Technologies. Morgan Kaufmann Pub-S, 2004.

[139] R. M. Martins, D. B. Coimbra, R. Minghim, and A. Telea, "Visual analysis of dimensionality reduction quality for parameterized projections," *Computers & Graphics*, vol. 41, pp. 26–42, 2014.

[140] J. G. S. Paiva, W. R. Schwartz, H. Pedrini, and R. Minghim, "Semi-supervised dimensionality reduction based on partial least squares for visual analysis of high dimensional data," *Eurographics Conference on Visualization (EuroVis)*, vol. 31, no. 3pt4, pp. 1345–1354, 2012.

[141] D. F. Andrews, "Plots of high-dimensional data," *Biometrics*, vol. 28, no. 1, pp. pp. 125–136, 1972.

[142] A. Inselberg, "The plane with parallel coordinates," *The Visual Computer*, vol. 1, pp. 69–91, 1985.

[143] F. Paulovich and R. Minghim, "Hipp: A novel hierarchical point placement strategy and its application to the exploration of document collections," *IEEE Transactions on Visualization and Computer Graphics*, vol. 14, no. 6, pp. 1229 –1236, 2008.

[144] D. Engel, L. Hüttenberger, and B. Hamann, "A Survey of Dimension Reduction Methods for High-dimensional Data Analysis and Visualization," *Visualization of Large and Unstructured Data Sets: Applications in Geospatial Planning, Modeling and Engineering - Proceedings of IRTG 1131 Workshop 2011*, vol. 27, pp. 135–149, 2012.

[145] F. Leliaert, D. Smith, H. Moreau, M. Herron, H. Verbruggen, C. Delwiche, and O. De Clerck, "Phylogeny and molecular evolution of the green algae," *Critical Reviews in Plant Sciences*, vol. 31, no. 1, pp. 1–46, 2012.

[146] J.-H. Choi, H.-Y. Jung, H.-S. Kim, and H.-G. Cho, "Phylodraw: a phylogenetic tree drawing system," *Bioinformatics*, pp. 1056–1058, 2000.

[147] 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.

[148] M. Hess, S. Bremm, S. Weissgraeber, K. Hamacher, M. Goesele, J. Wiemeyer, and T. von Landesberger, "Visual exploration of parameter influence on phylogenetic trees," *IEEE Computer Graphics and Applications*, vol. 34, no. 2, pp. 48–56, 2014.

[149] I. Elias and J. Lagergren, "Fast neighbor joining," *32nd International Colloquium on Automata, Languages and Programming*, vol. 3580, pp. 1263–1274, July 2005.

[150] M. Simonsen, T. Mailund, and C. N. Pedersen, "Rapid neighbour-joining," *Workshop on Algorithms in Bioinformatics*, vol. 3, pp. 113–122, September 2008.

[151] F. Stahl, B. Gabrys, M. M. Gaber, and M. Berendsen, "An overview of interactive visual data mining techniques for knowledge discovery," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 3, no. 4, pp. 239–256, 2013.

[152] N. Elmqvist, J. Stasko, and P. Tsigas, "Datameadow: a visual canvas for analysis of large-scale multivariate data," *Information Visualization*, vol. 7, no. 1, pp. 18–33, 2008.

[153] D. Keim, F. Mansmann, J. Schneidewind, J. Thomas, and H. Ziegler, "Visual analytics: Scope and challenges," *Visual Data Mining*, vol. 4404, pp. 76–90, 2008.

[154] S. Rüger, "Putting the user in the loop: visual resource discovery," *3rd International Conference on Adaptive Multimedia Retrieval: user, context, and feedback*, pp. 1–18, 2006.

[155] J. Poco, D. M. Eler, F. V. Paulovich, and R. Minghim, "Employing 2d projections for fast visual exploration of large fiber tracking data," in *Computer Graphics Forum*, vol. 31, no. 3pt2. Wiley Online Library, 2012, pp. 1075–1084.

[156] R. Etemadpour, R. Motta, J. G. de Souza Paiva, R. Minghim, M. C. F. de Oliveira, and L. Linsen, "Perception-based evaluation of projection methods for multidimensional data visualization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 21, no. 1, pp. 81–94, 2015.

[157] T. G. Dietterich, "An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization," *Machine Learning*, vol. 40, no. 2, pp. 139–157, 2000.

[158] T. Kohonen and P. Somervuo, "Self-organizing maps of symbol strings," *Neurocomputing*, vol. 21, no. 1, pp. 19–30, 1998.

[159] B. Settles, "Active learning literature survey," *University of Wisconsin, Madison*, vol. 52, no. 55-66, p. 11, 2010.

[160] G. Xuan, W. Zhang, and P. Chai, "Em algorithms of gaussian mixture model and hidden markov model," *International Conference on Image Processing*, vol. 1, pp. 145–148, 2001.

[161] X. Yuan, D. Ren, Z. Wang, and C. Guo, "Dimension projection matrix/tree: Interactive subspace visual exploration and analysis of high dimensional data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2625–2633, 2013.

[162] F. Zhou, J. Li, W. Huang, Y. Zhao, X. Yuan, X. Liang, and Y. Shi, "Dimension reconstruction for visual exploration of subspace clusters in high-dimensional data," *IEEE Pacific Visualization Symposium (PacificVis)*, vol. 1, pp. 128–135, 2016.

[163] B. Mory and R. Ardon, "Fuzzy region competition: a convex two-phase segmentation framework," *International Conference on Scale Space and Variational Methods in Computer Vision*, pp. 214–226, 2007.

[164] R. Malladi, J. A. Sethian, and B. C. Vemuri, "Shape modeling with front propagation: A level set approach," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 2, pp. 158–175, 1995.

[165] V. Caselles, R. Kimmel, and G. Sapiro, "Geodesic active contours," *International Journal of Computer Vision*, vol. 22, pp. 61–79, 1997.

[166] X. Bresson, S. Esedoglu, P. Vandergheynst, J.-P. Thiran, and S. Osher, "Fast global minimization of the active contour/snake model," *Journal of Mathematical Imaging and Vision*, vol. 28, no. 2, pp. 151–167, 2007.

[167] K. Zhang, L. Zhang, H. Song, and W. Zhou, "Active contours with selective local or global segmentation: a new formulation and level set method," *Image and Vision computing*, vol. 28, no. 4, pp. 668–676, 2010.

[168] T. F. Chan and L. A. Vese, "Active contours without edges," *IEEE Transactions on Image Processing*, vol. 10, no. 2, pp. 266–277, 2001.

[169] M. Rousson and R. Deriche, "A variational framework for active and adaptative segmentation of vector valued images," *Workshop on Motion and Video Computing*, pp. 56–61, 2002.

[170] S. C. Zhu and A. L. Yuille, "Region competition: unifying snakes, region growing, energy/bayes/mdl for multi-band image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, pp. 884–900, 1996.

[171] N. Paragios and R. Deriche, "Geodesic active regions and level set methods for supervised texture segmentation," *International Journal of Computer Vision*, vol. 46, no. 3, pp. 223–247, 2002.

[172] L. Gupta and T. Sortrakul, "A gaussian-mixture-based image segmentation algorithm," *Pattern Recognition*, vol. 31, no. 3, pp. 315–325, 1998.

[173] S. Lou, X. Jiang, and P. J. Scott, "Applications of morphological operations in surface metrology and dimensional metrology," *Journal of Physics: Conference Series*, vol. 483, no. 1, p. 012020, 2014.

[174] R. C. Gonzalez, *Digital image processing*. Pearson Education India, 2009.

[175] S. R. Sternberg, "Grayscale morphology," *Computer Vision, Graphics, and Image Processing*, vol. 35, no. 3, pp. 333 – 355, 1986, special Section on Mathematical Morphology.

[176] F. Mokhtarian and R. Suomela, "Robust image corner detection through curvature scale space," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 12, pp. 1376–1381, 1998.

[177] F. Ge, S. Wang, and T. Liu, "New benchmark for image segmentation evaluation," *Journal of Electronic Imaging*, vol. 16, no. 3, pp. 0 330 111–03 301 116, 2007.

[178] S. Alpert, M. Galun, A. Brandt, and R. Basri, "Image segmentation by probabilistic bottom-up aggregation and cue integration," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 2, pp. 315–327, 2012.

[179] H. Freeman, "On the encoding of arbitrary geometric configurations," *IRE Transactions on Electronic Computers*, no. 2, pp. 260–268, 1961.

[180] R. Vaddi, L. Boggavarapu, H. Vankayalapati, and K. Anne, "Contour detection using freeman chain code and approximation methods for the real time object detection," *Asian Journal of Computer Science & Information Technology*, vol. 1, no. 1, 2013.

[181] G. V. Pedrosa, M. A. Batista, and C. A. Z. Barcelos, "Image feature descriptor based on shape salience points," *Neurocomputing*, vol. 120, pp. 156 – 163, 2013.

[182] H. Blum and R. N. Nagel, "Shape description using weighted symmetric axis features," *Pattern Recognition*, vol. 10, no. 3, pp. 167–180, 1978.

[183] D. Shaked and A. M. Bruckstein, "Pruning medial axes," *Computer Vision and Image Understanding*, vol. 69, no. 2, pp. 156 – 169, 1998.

[184] M. Spitzner and R. Gonzalez, "Shape peeling for improved image skeleton stability," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1508–1512, 2015.

[185] A. S. Montero and J. Lang, "Skeleton pruning by contour approximation and the integer medial axis transform," *Computers & Graphics*, vol. 36, no. 5, pp. 477–487, 2012.

[186] L. d. F. D. Costa and R. M. Cesar Jr, *Shape analysis and classification: theory and practice*. CRC Press, Inc., 2000.

[187] K. R. Castleman, "Digital imaging processing," *Prentice Hall*, 1996.

[188] D. Chaudhuri and A. Samal, "A simple method for fitting of bounding rectangle to closed regions," *Pattern Recognition*, vol. 40, no. 7, pp. 1981 – 1989, 2007.

[189] J. Bendat and S. Sherman, "Monotone and convex operator functions," *Transactions of the American Mathematical Society*, vol. 79, no. 1, pp. 58–71, 1955.

[190] C. Chang, W. Liu, and H. Zhang, "Image retrieval based on region shape similarity," *Photonics West - Electronic Imaging*, vol. 1, pp. 31–38, 2001.

[191] S. Promdaen, P. Wattuya, and N. Sanevas, "Automated microalgae image classification," *Proc Computer Science*, vol. 29, pp. 1981–1992, 2014.

[192] M.-K. Hu, "Visual pattern recognition by moment invariants," *IRE Transactions on Information Theory*, vol. 8, no. 2, pp. 179–187, 1962.

[193] S. Abbasi, F. Mokhtarian, and J. Kittler, "Curvature scale space image in shape similarity retrieval," *Multimedia systems*, vol. 7, no. 6, pp. 467–476, 1999.

[194] T. Dharani and I. L. Aroquiaraj, "A survey on content based image retrieval," *International Conference on Pattern Recognition, Informatics and Mobile Engineering (PRIME)*, vol. 1, pp. 485–490, 2013.

[195] T. Sebastian, P. Klein, and B. Kimia, "Recognition of shapes by editing shock graphs," *IEEE International Conference on Computer Vision*, p. 755, 2001.

[196] T. Sikora, "The mpeg-7 visual standard for content description-an overview," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 11, no. 6, pp. 696–702, 2001.

[197] S. Jeannin and M. Bober, "Description of core experiments for mpeg-7 motion/shape," *MPEG-7, ISO/IEC/JTC1/SC29/WG11/MPEG99 N*, vol. 2690, 1999.

[198] H. Ling and D. W. Jacobs, "Shape classification using the inner-distance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 2, pp. 286–299, 2007.

[199] S. C, "Improving image retrieval using combined features of hough transform and zernike moments," *Optics and Lasers in Engineering*, vol. 49, no. 12, pp. 1384 – 1396, 2011.

[200] E. Walia, A. Goyal, and Y. Brar, "Zernike moments and ldp-weighted patches for content-based image retrieval," *Signal, Image and Video Processing*, vol. 8, no. 3, pp. 577–594, 2014.

[201] N. Laiche, S. Larabi, F. Ladraa, and A. Khadraoui, "Curve normalization for shape retrieval," *Signal Processing: Image Communication*, vol. 29, no. 4, pp. 556–571, 2014.

[202] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 4, pp. 509–522, 2002.

[203] N. Arica and F. T. Y. Vural, "Bas: a perceptual shape descriptor based on the beam angle statistics," *Pattern Recognition Letters*, vol. 24, no. 9, pp. 1627–1639, 2003.

[204] X. Bai, B. Wang, C. Yao, W. Liu, and Z. Tu, "Co-transduction for shape retrieval," *IEEE Transactions on Image Processing*, vol. 21, no. 5, pp. 2747–2757, 2012.

[205] M. Reif, F. Shafait, M. Goldstein, T. Breuel, and A. Dengel, "Automatic classifier selection for non-experts," *Pattern Analysis and Applications*, vol. 17, no. 1, pp. 83–96, 2014.

[206] M. Girolami, "Mercer kernel-based clustering in feature space," *IEEE Transactions on Neural Networks*, vol. 13, no. 3, pp. 780–784, 2002.

[207] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[208] E. Çomak and A. Arslan, "A biomedical decision support system using ls-svm classifier with an efficient and new parameter regularization procedure for diagnosis of heart valve diseases," *Journal of Medical Systems*, vol. 36, no. 2, pp. 549–556, 2012.

[209] D. M. Powers, "Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation," *Flinders University*, 2011.