Facultade de Informática
UNIVERSIDADE DA CORUÑA

TRABALLO FIN DE GRAO
GRAO EN ENXEÑARÍA INFORMÁTICA
MENCIÓN EN COMPUTACIÓN

Automatic system for the detection and recognition of phytoplankton in digital microscope imaging

Estudante: David Rivas Villar
Dirección: Jorge Novo Buján
Dirección: José Rouco Maseda

A Coruña, agosto de 2019.

To everyone who has always supported me

Acknowledgements

I would like to thank everybody that has helped me along the degree and in this very project. In first place, my directors Jorge Novo Buján and José Rouco Maseda for providing me with this opportunity to work with them and guiding me along the development of this project. I am also really thankful to Rafa that solved my many doubts about plankton and the biological part of this work. I would also like to thank my colleagues from the VARPA group for creating a comfortable work environment and being always down to help.

I am also really grateful to all my friends and companions that have always lent a hand and tolerated me when I was the most unbearable. A special thank you to my family and girlfriend Gemma who have all been very supportive of me and have made possible that I am writing this very lines.

Abstract

The quality of water can be compromised by the proliferation of toxic species of phytoplankton. When these blooms occur in rivers and reservoirs used for the water supply, this event can have a negative impact on human health. Currently, to determine the existence of risk, experts rudimentarily monitor phytoplankton populations by sampling and analysing the water. This analysis consits on the identification of dangerous species and its biologic volume. All in all, this process is long and tedious when the amount of samples that need to be analysed in order to obtain a quality and representative measure is taken into account, which also needs to be carried out periodically for each water source. The taxonomic process requires broad experience and training for the personnel involved. The automation of these tasks is highly desirable as it would free the experts from part of the work at the same time as that eliminates subjective factors that may impact in the overall quality of the process.

In this work the intention is to help experts starting from images obtained directly from a conventional microscope, differentiating it from other similar works in the state of the art that use specific hardware. Computer vision techniques will be used to detect candidate individuals and artificial intelligence methods to recognise relevant phytoplankton species, that is, the toxic ones, distinguishing them from the rest of objects in the images like, for example, inorganic materials. Finally, the phytoplankton organisms will be classified to obtain a metric that counts the dangerous ones and so be able to analyse the quality of the water.

Resumo

A calidade da auga pode verse ameazada pola proliferación de especies tóxicas de fitoplancto. Cando estas proliferacións ocorren en ríos e encoros utilizados na subministración de auga potable este feito pode ter impactos negativos na saúde humana. Actualmente, para determinar a existencia de risco, os expertos monitorizan, de forma rudimentaria, as poboacións de fitoplancto mediante a recolección de mostras e a súa correspondente análise. Esta análise consiste na identificación das especies perigosas e o rexistro do seu volume biolóxico. Con todo, este proceso resulta longo e tedioso se se ten en conta a cantidade de mostras a analizar para poder ofrecer unhas métricas fiables e representativas, as cales se deben realizar periodicamente para cada unha das fontes de auga destinadas ao consumo. Así mesmo, o proceso taxonómico require unha ampla experiencia e formación específica do persoal involucrado. A automatización destas tarefas é moi desexable xa que libera aos expertos de parte do traballo, á vez que evita factores subxectivos que poidan influír na calidade global do proceso.

Neste traballo preténdese axudar aos expertos partindo de imaxes obtidas directamente do microscopioconvencional, diferenciándoo de traballos similares do estado do arte que requiren hardware específico. Empregaranse técnicas de procesado de imaxe e visión artificial para detectar individuos candidatos e técnicas de intelixencia artificial para recoñecer as especies de fitoplancto relevantes, é dicir, as tóxicas, distinguíndoas do resto de obxectos nas imaxes, como, por exemplo, materiais inertes ou inorgánicos. Por último, os microogranismos de fitoplancto son clasificados para obter unha métrica que contabilice os perigosos e poder, así, analizar a calidade da auga.

Keywords:

- Computer vision
- Phytoplankton
- Image segmentation
- Taxonomic classification
- Bag of visual words with Gabor filters
- Machine learning
- Cyanobacteria

Palabras chave:

- Visión por computador
- Fitoplancto
- Segmentación de imaxe
- Clasificación taxonómica
- Bolsa de palabras visuais con filtros de Gabor
- Aprendizaxe máquina
- Cianobacterias

Contents

List of Figures

List of Tables

Chapter 1
Introduction

In this chapter the work is introduced providing the context and the motivation that culminated with the realisation of the project. The objectives and the different parts of this endeavour are explained to provide the basis for the project.

1.1 Motivation

At present, the studies about plankton and other types of microbes require a great deal of effort for coordinating different teams and following standardised procedures. Biologists, limnologists and all sorts of scientists and researchers need to collect samples of the water *in situ* to be analysed, be it a river, a lake, on the shore, etc. This alone can require a lot of time but it is only the beginning of the whole process. After picking up samples, they need to be treated and preserved to be observed in a suitable place with a microscope. Once the samples are under the microscope, they are observed and digital images are captured to be later analysed or just for archival purposes. The analysis varies depending on the type of water contained in the samples but, more importantly, depending on the objective of the study. While some studies focus on quantifying the biomass of certain phytoplankton species in order to determine the water conditions, other are more complex and require researchers to classify all the organisms present, and measure them in various ways to study, for example, seasonal changes in size.

This process is complex and comes with a huge workload for specialists. Furthermore, the variations between the criteria of different experts might hurt the quality of analysis, as the classification and quantification of specimens is a subjective task up to the point of providing varying results for two analysis performed by the same expert. With an automatic tool this bias could be mitigated or completely eliminated, as it could help researchers to do part of the tedious work for them in an objective and repeatable way. This project is focused on providing computer-aided tools to help researches to simplify the tasks that, coincidentally, are the most likely be done periodically: the tasks included in a water potability test like individual recognition and classification of them into species. As different blooms of phytoplankton occur on dams the water could become dangerous for human consumption if the toxins segregated by the phytoplankton reach a certain point. The objective of this work is to detect phytoplankton in digital microscopy images, classify said specimens and determine if they are in any taxonomic group that could lead to health problems in humans. These tasks are oriented to, ultimately, provide an objective and repeatable way to quantify the blooms of the dangerous species

1.2 Objectives

This project can be divided in two main parts or objectives. The first is the development of a system capable of identifying candidate phytoplankton specimens in digital images of samples created using a microscope. This system should be also able to differentiate single specimens from colonies, as the latter is often the way many species of phytoplankton appear in the wild.

The second objective is to classify these candidate specimens, first, into two broad classes: phytoplankton and non-phytoplankton. This is done to reduce the rate of false positives, as the samples are real world ones and they usually contain all sorts of non-relevant objects like detritus, cloth fibres, plastics, sand particles, etc., as well as other plankton species like zooplankton, that are not relevant in this type of studies. Once the separation between what is interesting and what is not is achieved, the next step is to recognise the target species of phytoplankton. The main point in this step is to identify the presence of dangerous species in water and monitor their amount, to keep the toxicity of the water under control.

1.3 Structure of the dissertation

This dissertation is organised in seven chapters which include this introduction, a contextualisation and description of the domain followed by the planning for the project. After those, the fourth chapter is focused on the automatic segmentation of the images and the detection of the target object candidates. Chapter five describes the classification methodology used to first, identify phytoplankton candidates and, then, recognise the target species. Finally, chapters six and seven detail the conclusions and future work of this project, respectively. Furthermore, this work includes two appendices with extra materials that could not be fit into the main dissertation due to space concerns.

- **Chapter 1: Introduction.** This first chapter describes the motivation of this project as well as the objectives for the work. It ends with this description of the dissertation's structure.

- **Chapter 2: Description of the domain.** The second chapter details the context in which the project takes place, explaining the fundamentals of plankton and its importance in any water-related ecosystems. The collection, treatment and digitisation of the samples is detailed for the type of data used in this work. To finalise this chapter a section describing the state of the art is included.
- **Chapter 3: Planning.** In this chapter the planning made at the beginning of the project is described as well as any possible deviation that occurred during the course of the development. Tasks, resources needed and cost estimations are detailed in this section.
- **Chapter 4: Image segmentation.** In this chapter of the dissertation the first step of the process is described. Here the image is segmented and the possible candidate specimens are detected. The obtained and results are analysed following a specific methodology, that is also described in this chapter.
- **Chapter 5: Classification.** The obtained specimen candidates in chapter 4 are classified in this chapter. First, into phytoplankton and spurious elements, and then into dangerous cyanobacteria and other phytoplankton species. This chapter details the testing of several distinct features of the specimens as well as several classifiers with the intention of finding out the best performing ones for this scope and each one of the classifications created.
- **Chapter 6: Conclusions.** All the conclusions of this project are laid out according to the methods used and the results obtained.
- **Chapter 7: Future work.** This last chapter includes all the work that could expand and possibly improve the current project.
- **Appendix A: Theoretical explanations.** Contains theoretical explanations and other information about certain methods that due to their length could not be included in the main dissertation.
- **Appendix B: Extra figures.** Contains images that could not be included in the main dissertation due to space concerns.
- **Bibliography.** List of all bibliographic resources used along the development of this project

Chapter 2

Description of the domain

This chapter explains every concept needed to fully understand this work, providing context on the currently used techniques to carry out phytoplankton analysis as well as explaining the basics of what is phytoplankton and why is it so relevant to the environment and the human beings.

First the basics of what phytoplankton is are described, detailing its taxonomic nature as well as commenting its relevance in different areas. Then a state of the art revision is provided digging into the existent methods to analysed and quantify phytoplankton that exists highlighting their strengths and weakness. Finally a detailed explanation of the differences wit the proposed work is provided.

2.1 Phytoplankton

Plankton studies date way back to 1828 when the amateur naturalist J. Vaughan Thompson first towed his fine-meshed net with the intention of capturing plankton. The word plankton was defined later, in 1887, by professor Victor Hensen.

Phytoplankton is, together with zooplankton and other less relevant groups, a subgroup of plankton. Plankton encompasses all the organisms that, by their locomotion ability, are incapable of swimming and thus are transported by the water currents [1]. This fact is also reflected in the name of the group as it comes from the Greek root *planktos* which means ”that which is passively drifting or wandering”. Despite this definition, some species of phytoplankton can adjust their position in the water rising towards the surface or drowning towards the bottom.

The distinction between phytoplankton and zooplankton depends upon whether the organism is considered a plant or an animal, respectively. Using a different perspective, plankton are of usually small size. However, that is not always the case, as some types of zooplankton can measure up to meters in diameter [1]. This is relevant as it is also common to classify plankton by size, which reveals several groups like megaplankton, macroplankton, picoplankton, etc.

Finally, plankton can also be classified by the time an organism stays in said group. That is, if they spend their whole life as plankton, they are said to be holoplankton. However, if they only spend part of their lives as plankton, like for example in larval stage, they belong to meroplankton.

Phytoplankton is divided in several groups like diatoms, dinoflagellates or cyanobacteria. Cyanobacteria, also know as blue-green algae or blue-green bacteria, contain both some of the largest and smallest species of phytoplankton.

To measure or analyse phytoplankton it is usual to take samples from the water body that is going to be analysed and take them to a laboratory. There, the sample is processed mainly into two different measures, standing stock and biomass. Standing stock describes the amount of organisms per volume of water at the moment of sampling. This is usually done by microscopy count of microorganisms in each sample. This metric is far from ideal as phytoplankton vary a lot in size so it is more appropriate to measure the amount of these organisms in terms of biomass. Biomass is the total weight of specimens per volume of water. As the individuals can not be possibly weighted manually, the organism count is multiplied by the average weight of the species in order to get a representative estimate. Another way to estimate the amount of phytoplankton in water is to use the fact that all species contain chlorophyll, that can be easily measured with a fluorometer. This yields results in amount of chlorophyll by volume of water.

The great majority of plants in oceans are of planktonic nature, many of which are of microscopic size. As the phytoplankton are the dominant plants in the ocean their importance in the marine food chain can not be underestimated.

Phytoplankton need both light and nutrients to survive so they tend to develop in places with the right amount of both. A good example is river mouths, regardless of them flowing into the sea, a lake or a dam. At these points, rivers come loaded with nutrients which motivate the development of big populations of phytoplankton. These favourable conditions motivate blooms, which are waves of increased development of phytoplankton in which their population grows rapidly. Blooms are well known by the effects of red tides which is, a type of bloom caused by the increased amount of dinoflagellate. These blooms may cause fish death by oxygen depletion but, more importantly, certain species that are prone to bloom can produce a series of neurotoxins. These toxins are often consumed by humans in shellfish, which could lead to death. Other common blooming group are the cyanobacteria that can also create dangerous toxins for humans.

Cyanobacteria are perhaps some of the most important groups of phytoplankton [2]. Due to their production of oxygen through photosynthesis they played a key role in the early oxygenation of Earth, and continue to do so. This type of plankton can be found in almost any habitat in the world, be it terrestrial or aquatic. Some of them produce toxins and can become dangerous when their population grows too much in a bloom. Cyanobacteria can develop in low light environments due to their accessory pigments. Some of them have the capacity to rise up to the surface using gas vesicles which allows them to form scum granting easier access to solar light. These and many other adaptations have made this group of phytoplankton particularly prolific around the world [3].

2.1.1 Importance, problems and remarkable aspects of phytoplankton

Phytoplankton is the basis of the food chain in all the water bodies in the world, either being freshwater such as rivers or saltwater like oceans. Trough photosynthesis, they are able to create high-energy organic materials from carbon dioxide and water plus inorganic nutrients. Being the dominant plants in the oceans they can be equated to rainforest and other fertile habitats that everybody can identify as really productive and important. Phytoplankton is frequently overlooked as one of the main sources of oxygen on earth. Trough the previously mentioned photosynthesis phytoplankton is able to produce about 50% of all the oxygen on earth [4]. They also act as carbon sink, since when they die part of the carbon that their bodies accumulate gets buried and, with time, can form petroleum or coal.

However, phytoplankton also poses a risk to the creatures that consume the water where it lives in, as some species are potentially poisonous. Like the previously mentioned dinoflagellate, some cyanobacteria are known to produce a variety of potent toxins, commonly known as cyanotoxins, as well as some compounds that alter the taste and/or odour of the water. The most common cyanotoxins are microcystins, nodularins, anatoxins, cylindrospermopsin, etc, which have different effects on humans, ranging from gastroenteritis to liver damage and cancer [5]. These cyanotoxins are secondary biologically active metabolites that are derived from pigments that build up in the cellular interior. They present diverse chemical structures and can not be eliminated trough conventional water purifying systems, which aggravates even further their dangerous effects on human health [6]. The production of this toxins is linked to the blooming phenomenon, that allows these organisms to drastically increase their concentration in water. This important increase of biomass leads to levels of cyanotoxins in water that can be potentially dangerous [6, 7, 8], not only for drinking, but also for recreational use like swimming or water sports. Adverse effects have also been linked to aerosol inhalation [9], which can be produced during water sports without the need to even be soaked in water. It has been found that nasal exposition is comparable, in terms of toxicity, to that of a intraperitoneal injection.

Short but intense exposures (that is, being exposed to highly contaminated water for a short period of time) can result in the previously described adverse conditions. However it is often overlooked that the chronic low dosages of toxins that are, sometimes, present on normal water can also be harmful. Repeated exposures to microcystins, even on small quantities, can create long lasting problems for people. Long-term effects have been observed promoting liver damage and even liver cancer.

The increased frequency and intensity of cyanobacterial blooms have been linked with climate change as it appears to promote these blooms in surface waters all over the globe due to the increase in average temperature [10]. While some studies have been made on the matter, it is still largely unknown what exactly cause these type of blooms. Europe-wide studies have highlighted eutrophication levels as one of the factors of increased cyanobacterial proliferation. This remarks the necessity of continued and reliable studies over dams and other sources of water. These control studies are not only necessary to asses the safety for human consumption but also for cattle and crop irrigation. This is because human health could also be compromised by eating meat or plants that were in contact with poisonous water [5]. This water also has effects over the said animals and plants, potentially killing or harming them.

Even if many advancements have been made on the matter, it is still very difficult to predict population dynamics more than a few days in advance. Population and toxicity of blooms are mainly influenced by different environmental pressures. Environmental pressures are all the factors that can change a sub-aquatic ecosystem like water circulation, human intervention, climate variability, etc. These pressures make the blooms on similar areas take completely different trajectories on their development. Thus, for the same ecosystem large variations can be found from year to year, which highlights even further the volatility and unpredictability of these blooms. An appropriate water sampling and control remains the best strategy, and it must be adapted to the kind of monitoring needed as well as to the local conditions.

Harmful effects of cyanobacterial blooms have motivated authorities around the globe to adopt strategies to handle this type of events. The World Health Organisation (WHO) has also issued a series of guidelines that most of these strategies are based on. These guidelines suggest using two alert levels depending on the level of cyanobacteria in drinking water sources, which are limited by concentrations of 2,000 and 100,000 cells/mL. At the European level, measures have been taken to monitor these blooms and toxins [11, 12] which have been already implemented in the Spanish legislation [6]. Specifically in Galicia, the Xunta monitors cyanobacteria in both bathing waters and potable water supply [13, 14]. This monitoring is also done by the companies running the reservoirs.

In Spain it is estimated that at least 25% of all lakes are in danger of developing at least one toxic episode a year. Galicia stands out in Spain as, due to its mainly granitical lithology, which makes it is more likely that lakes develop cyanobacterial blooms [15]. Several of these toxic blooms have already been reported and have been acknowledged by the authorities [16, 17, 18, 19]. An example of one of these blooms [19] can be seen in figure 2.1.

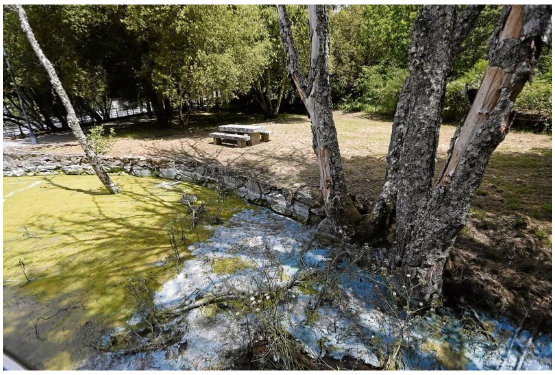


Figure 2.1: Example of cyanobacterial bloom in As Conchas reservoir

Bloom monitoring is a really important task that should be carried out frequently. However, the procedures used to study such dangerous events are still mostly relying on manual means. Recent studies on lakes all over Europe [20, 21] had to make special efforts in order to minimise subjectivity, create a uniform analysis data. A special, concrete protocol was not enough for this studies as differences between experts' judgement could prove too big. Instead they required that the local team in charge of gathering the samples from a lake to send them frozen to be analysed in a central laboratory. There, all the samples were analysed by the same team, which allowed them to gather more homogeneous data. Seeing this complex procedure it easily inferable that the analysis made by different people are vastly different. Apart from this bias there are other factors that influence the quality of analyses like boredom, tiredness, etc. Furthermore, the biomass analysis is also influenced by changes in the preservation method and duration. These factors can introduce up to a 40% of variation in results as they alter the cyanobacterial morphology [5]. These changes are so drastic that they could mean that dangerous water that may be mislabelled as safe water to be distributed to homes and cattle, used for swimming, etc. which poses a great threat to public health.

2.2 State of the art

Currently, the phytoplankton analyses to assess water potability are, still being performed manually. This requires an expert to collect water samples from the reservoir that is being studied. Then each water sample is put under the microscope and the expert identifies plankton organisms. This are then taxonomically classified into species and, finally, a count of the dangerous ones is made. The count is the metric used to determine the quality of the water. This method is prohibitively costly in terms of work as it may require two to five days to analyse a batch of samples [5], which in the case of critical applications such as potability analyses may be too much. This is the reason why, along the years, many innovations have emerged trying to alleviate the workload of experts or at least provide them with some help.

In order to provide a revision of the works aiming to automate parts of the process behind phytoplankton analyses, therein two main stages can be identified: image acquisition and image analysis. Image acquisition includes the gathering of water samples, their processing to analyse them with a microscope and the digitisation into images. On the other end image analysis includes the computer process that allows to segment the images into regions containing each plankton specimen, classify them into known species, or other suitable classes. It should be noted that every work should clearly state how these two stages are performed despite being mainly focused on innovation in one of them.

2.2.1 Image acquisition

Image acquisition focuses on the gathering of the different water samples and the process which involves transforming them into images. Currently this process can be divided mainly into manual and automatic procedures.

Manual image acquisition

The most common process for gathering phytoplankton images is usually through the normal operation of a microscope. The samples used in this method are usually gathered by the experts in a lake or reservoir using a boat and specific instruments that allow to take samples of water at different depths. This samples should be taken periodically with the period changing depending on the type of analysis required. The sampled water is then concentrated and preserved using different chemicals like glutaraldehyde. The solution used to preserve them can alter the morphology of the cells and thus the biomass [5].

Once the samples are fixated and processed into a microscope slide the expert may proceed to examine them. The microscope focus and movement is operated manually by the expert, who can perform the analysis and counting doing this operation. However it is also common that these images are digitised and saved for later processing or revision of the results.

The whole process takes a considerable amount of time for the experts and may have problems of repeatability due to the sampling, preserving or imaging process if no clear procedure is developed.

Usually the expert takes samples of the water and by putting them under the microscope the digital images are obtained. The process is not equal everywhere but it is common is the fact that this process takes a fairly long time and deal of effort.

Other kind of devices use fluorometric probes. Fluorometric or fluorescence probes are a popular method for analysing phytoplankton colonies.They work by using the chlorophyll that cyanobacteria contain, this chlorophyll emits light in a certain wavelength when shone a certain light as well. This allows a simple spectrum analysis to determine the amount of cyanobacteria in a sample of water. This number can then be used to determine how safe is the water for consumption. While fluorometric analysis is not as precise as the microscope, it can provide valuable information requiring much less effort. Varying pigment cell content, cell size, cell agglomeration, water turbidity or temperature are among the many factors that can alter the result that this type of probes offer [5]. The main advantage of these methods is the fact that the amount of work is way lower than the more precise microscope imaging method. This highlights the need for a faster method that has less compromises than fluorescence analysis.

Fluorescence is often used as a complementary to other common imaging techniques (manual or automatic) since it makes identifying phytoplankton and more specifically cyanobacteria trivial [22, 23, 24, 25, 26, 27]. By shining a light to the samples the phytoplankton, phytoplankton organisms glow distinctly which allows simple image processing to detect and segment the images with ease. This is particularly useful since most of the confusion could come zooplankton or other biological particles that are similar in colour, size, texture, etc. They, however, do not shine with the fluorescence light applied which allows this systems to easily discard those otherwise complicated cases.

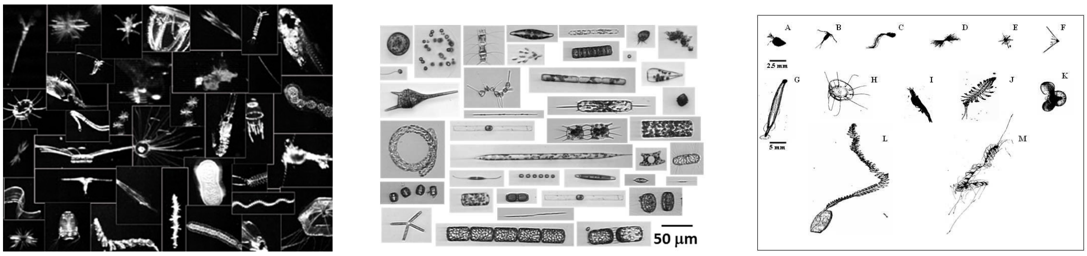
Automatic buoys that detect photosynthetic pigments in the water [28] this way have also been developed and are used to monitor the presence of cyanobacterium in the water.

Automatic image acquisition

The need for automation in the acquisition also motivated different types of hardware implementations to alleviate the work of specialists. The different implementations can be categorised into two types: underwater camera equipment and laboratory equipment. The first type is composed of a series of special cameras, generally attached to microscope that are able to take images of the phytoplankton in the water as they are moved with the help of a boat of other type of propulsion. Laboratory equipment, on the other hand, still requires the experts to collect the water and to deposit it on the machine. Both types of contraptions output images of the plankton they detect in the water, with the underwater camera type also solving the problem of water sample gathering.

The earliest example an underwater camera of sorts is Video Plankton Recorder (VPR) [29, 30]. This system is a video microscope, which allows to record microscopic organisms. Its output images are grayscale, being able to 60 images per second of particles down to 50 microns in size. This instrument is quite bulky, it needs to be towed by a boat to capture the images so it can not be deployed in many freshwater bodies.

A similar system is SIPPER or Shadowed Image Particle Profiling and Evaluation Recorder. SIPPER lets the water flow through a small tubular opening and takes pictures using a linescan camera. This results in binary images, that is, black and white. It does not need to be towed



(a) Video Plankton Recorder sample images

(b) FlowCytobot sample images

(c) SIPPER sample images

Figure 2.2: Sample images of different automatic systems

as it is integrated in an autonomous vehicle, but its size may prevent it from being used in shallow waters like in the previous case.

KRIP [31] was created to improve upon the VPR by downsizing it which would make it easier to tow and to able to deploy in more shallow waters while using similar image capturing techniques.

FlowCytobot [32, 33, 34] is another automatic system to perform acquisition. This system can output images in a higher quality than those described above, although still in grayscale. It is also way smaller which could allow it to be used in some freshwater environments.

All this systems allow to gather data automatically from the reservoir which significantly reduces the work of specialists. Their main drawbacks are their size, that could prevent them from deploying on shallow freshwater and their cost. They also do not save the water samples used in the images as they just photograph the water that flows through them.

The most relevant hardware tool of all belongs to the laboratory equipment. FlowCam [27, 35, 36, 37, 38, 39] does not gather the sample, an expert needs to do so. The trade-off is much higher image quality, nearing that of a microscope. The expert needs to introduce the water, previously sampled, in the machine and, without having to manually identify the organisms or do anything, the system produces an output which consists on images of the individual microorganisms. The main problem with this pieces of hardware is the price which ranges from $20k all the way to $100K.

Figures 2.2c, 2.2b, 2.2a and 2.3 show example images acquire using the SIPPER [40], FlowCytobot [34], VPR [41] and FlowCam [42], respectively. Looking at the resulting images from each of these systems it is easy to asses their difference. SIPPER gives the most basic results with binary images , both FlowCytobot and VPR output colourless images while FlowCam uses regular colour images. The FlowCam are thus the most similar to the focus of this work. Images from some of this tools are the ones normally used in the classification works. While most of the images and the derived works related to VPR and SIPPER are zooplankton, it is possible to use both for phytoplankton.

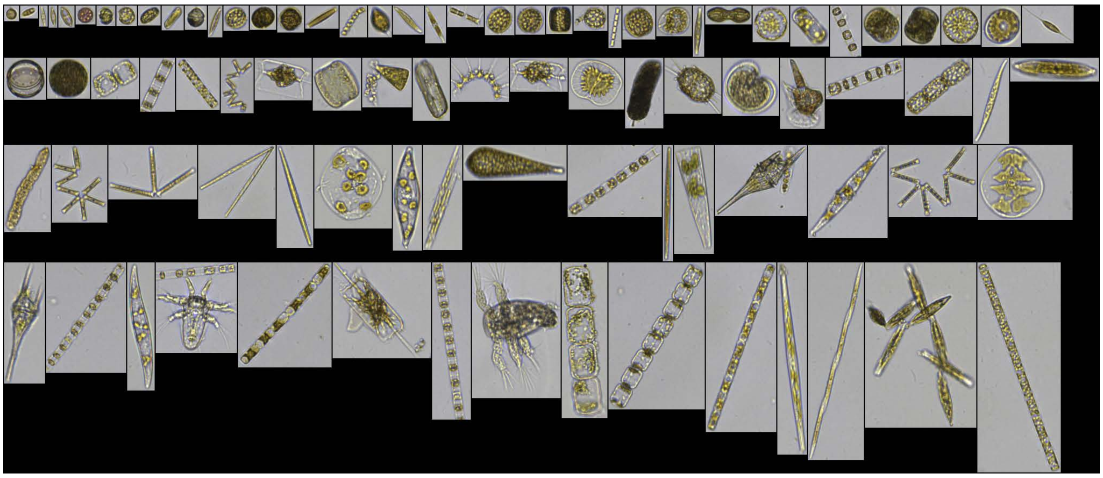


Figure 2.3: FlowCam sample images

2.2.2 Image analysis

Image analysis includes both organism detection and classification of said detections. Both of these tasks are the main focus of this work. Organism detection is the process that allows to separate all the candidate organisms in the image from the background. The classification, instead, allows to separate the candidate organisms by a certain criteria. For example, it can help differentiating between harmful and harmless specimens or simply identifying each of the species.

Currently the most common way of handling this process is, like previously, manually even if some automatic systems have been proposed along the years.

Manual

As previously stated the most common way of handling these type of analyses is manually through a microscope. In this case the expert puts the sample in a microscope slide and carefully looks at every individual. This offers advantages over using still images as it allows the expert to control the focus and the magnification which allows for better focusing and distinction between species. The expert must identify which objects in the sample are phytoplankton and then classify them taxonomically. In the case of a potability test the expert must also count the dangerous cyanobacterium to obtain a metric of biomass or cells per unit of volume in the sample to determine the quality of the water. It is also common that these microscopes are equipped with cameras which allows the biologists to take pictures of the sample for archival or revision purposes.

Automatic

Many of the previously mentioned automatic image acquisition systems only provide the segmented image, that is, each individual of plankton is enclosed on its own image, not the whole picture. This corresponds with the segmentation step. The algorithms used in this process have not been publicised as they are commercially sold together with the hardware, like the case of FlowCam.

In the work by Haiyong Zheng et al. [25], phytoplankton organisms are detected using an Information Gain (IG) saliency object detection [43] as well as a saliency detection by saturation. Saliency maps are a measure of the importance of each pixel in the image, it tries to detect the most salient and attention-grabbing objects in the image. Various metrics are employed in this process, saturation measures the difference in saturation of the objects in an image. The saturation is the colourfulness of an area judged in proportion to its brightness [44]. IG, on the other hand, essentially tries to quantify how much better a posterior predicts the data than a prior [45], that is, according to a baseline how another model can predict a result. If the new model can do it better than the baseline, it holds more information, in turn, if it is worse the selected features hold less information and thus are worse. Both of these results are then fused and binarised with Otsu's method. Due to the imaging process the used images presented spurious illumination on the corners of the images so they just discarded them as no organisms are usually in those places. The resulting map is then eroded and dilated through mathematical morphology to create markers for the watershed algorithm. Watershed is a segmentation algorithm that is usually based around starting its process from user-defined markers. The markers signal the algorithm the points where the flooding should start. This flooding is done attending to the main feature of the Watershed algorithm that treats the pixels intensity values as topography or elevation and acts like water flooding the basins attributed to different markers. The process in this work creates markers automatically, eliminating the need for a user to define them. The Watershed algorithm is applied to the gradient image obtained directly from the original colour image yielding the final segmentation between specimens and background.

In the work by Verikas et al. [46] the base images are those obtained through a microscope, similarly to the proposed work. This work is only focused the *Prorocentrum minimum* phytoplankton species. *P. minimum* specimens are round, and an ad-hoc technique is used to search for circles, similar to what a Hugh transform would do but using phase congruency. Circles too small or too big to be *P. minimum* are discarded and morphological opening is applied to the remaining ones with the intention of cleaning the results.

In the paper describing the KRIP system [31], mentioned above, a novel algorithm for image segmentation and plankton detection is provided. The microscope image is first preprocessed by demosaicing and re-sizing it. After that, it is applied a Sobel filter which would only detect in focus particles. Then the images are binarised with a empirically found fixed threshold value and, finally, a morphological process consisting in dilation and erosion is applied with the objective of reducing noise and imperfections. The labelling process selects the connected pixels as regions after the whole process and bounding boxes containing each ROI are extracted.

Regarding the classification of organisms, there are more works in the state of the art as they make use of sets of images obtained by the previously mentioned systems like SIPPER of FlowCam [27, 37, 39, 47]. Even if some of these works use the same set of images each one of them tries to test some features or systems not yet proved.

Classification of plankton organisms, specially phytoplankton is particularly difficult since the organisms are inherently deformable, may have different shapes depending on which axe the image is taken from and, in general, can present a lot of dimorphism between individuals while different species keep being relatively similar. Dimorphism between individuals may not only be in shape but also in colour or texture since depending on many variables such as temperature or age, organisms may contain more or less chlorophyll or any other component that gives them their distinctive features.

There are many classifiers for zooplankton however they are not usable for phytoplankton, not only because their features are different but because of their size [29, 48]. Phytoplankton is often much smaller which means the methods used to identify and classify zooplankton usually do not work at enough resolution to solve these tasks on phytoplankton. This increases the division between systems as those used in zooplankton can not be used for phytoplankton limiting their scope. This creates many different systems that can not be reused for other similar, albeit different, applications.

Many of the current classifier systems about phytoplankton are centred on marine species [46, 49, 50] since they are of special interest not only for the environment but also for food production as red tides can contaminate many seafood. Despite its importance freshwater has been left kind of untapped as many places still lack proper safety measures to monitor water.

If the differences between saltwater and freshwater as well as those between phytoplankton and zooplankton, are ignored there are a lot of different works in the state of the art where valuable information can be found. Many implementations follow each others path with similar findings and development.

Focusing just on image classification, there are a lot of works that have achieved good results with a wide variety of different classifiers. There are examples using Support Vector Machines (SVM) [51, 52, 53] and many other classifiers like k Nearest Neighbour (k-NN) [37, 37], Artificial Neural Networks (ANN) [49, 50, 54], Random Forest (RF) [24, 46, 55] and even some deep learning methods [39, 56, 57].

The features that the classifiers mentioned above use very much vary among the different work depending on the scope that they target. This is mainly due to the fact that some works target very specific scopes like looking for particular organisms, as e.g. spherical plankton and thus only look for spherical plankton features [46]. The most common features are geometric ones like major and minor axis length, perimeter, area, etc. Texture properties like Local Binary Patterns [58], Gabor filters [59], Hu moments [60], etc. are also a common source of features. Some works go a little bit further and try to detect some special features of the microorganisms they want to identify, like flagellums or other cellular structures [50]. Additionally, most of the works tend to create a very big vector of characteristics and narrow it down using some kind of selector which chooses the most representative features for their image set and the ones that are most useful to the chosen classification algorithm [61, 62, 63]. A concrete example of this can be seen in the work by Verikas et al. [46]. In this work classification is also made, 65 features were extracted including several of shape like area or perimeter and others of texture like intensity and standard deviation in Gabor filters. Gabor filters are used for texture analysis and, intuitively, check if certain frequency is localised in a specific direction in a certain region of the image. A committee classifier between Random Forest and SVM outperformed these two classifiers by themselves resulting an 94.9% of accuracy.

Most of these papers use some kind of already made database where plankton organisms have been processed into images where they are alone, without any debris or junk that may difficult classification. These ideal case scenarios are not representative of the real world, and, while they can look great on paper, the results of the system in a real scenario may be wildly different as the real world samples are far from optimal. Generally they contain trash, dead individuals, zooplankton, individual overlapping, etc. This makes the works that use database with selected images differ from a real world perspective.

2.3 Proposed work

This work, starting from microscopy images, detects, segments and classifies organisms. All the necessary work for potability testing and phytoplankton analysis is enclosed in a single project. The only possible things not in the scope of this work are sample collection and biomass estimation. Biomass estimation was deemed not necessary as the metrics for water quality can be in biomass per litre but also in dangerous cell count per millilitre, which is very similar to the output of a the species classifier.

The main objective of this work is to be deployable in any laboratory. Most of the already reviewed systems require expensive equipment that many places may not be able to afford. On the contrary the proposed work only needs a microscope that can be equipped with a camera, equipment readily available in most laboratories.

The proposed work also intends to segment real world images , that is, images with many real specimens per image. These are usually accompanied by trash and many other spurious elements. These particular features separate this work from the rest as the state of the art is mainly composed from works that select the images that they use.

The methods tested for classifying the species in this work are also innovative as they have not been used in any of the previous works. The features are based on a bag of visual words with Gabor filters.

2.3.1 Sample collection and treatment

The water samples used in this work were collected on the Doniños lake in Ferrol and all belong to the same batch. These samples were collected using a zodiac-type boat in the deepest part of the lake, that amounts for a maximum of 11 meters deep. The sampling of water was done for phytoplankton analysis as well as chemical analysis like nutrients, metals and anions or cations present in greater amounts. The water is collected using a van Dorn bottle of 5 litters de Walt at four different depths: 0, 3, 6 and 9 metres. This process is done bimonthly since 2012. Together with this samples data was gathered in situ using a Secchi disk and physicochemical profiles through a multiparametric probe.

The volume of water used for the phytoplankton analysis was originally of 0.5 litres. It was, however, concentrated using a filter of 0.45 µm of nitrocellulose acetate. The resulting volume was 50 mL to which glutaraldehyde was added as means of preservation in a solution of 10% of concentration.

These samples then were processed into microscope slides and examined under an optic microscope Nikon Eclipse E600 equipped with an objective E-Plan 40x (N.A. 0.65) and a optical differential interference contrast (Nomarski) that allows for better contrast in these transparent samples. Finally the microscopy photos where taken using a digital camera AxioCam ICc5 Zeiss.

As the expert can control both focus and magnification of the images several options could have been used. Originally a 40x magnification was presented but as only a specimen was to be seen per image it would still require lots of work by the experts. It was then deemed that the best approach was to reduce magnification which would allow experts to work faster as less magnification means more of the sample can be in each photo. The selected magnification was 10x which encompasses many specimens per image.

2.3.2 Resulting images

Resulting images from the previous process have not been further processed or cleaned up in any way.

These images are quite different to those used by the methods reviewed in section 2.2 as they have specific features like colour or a higher degree of quality in terms of resolution. The images are also not selected or picked in any way, that is, all the images used are results of real world imaging without any discards. This mainly means that these images have a high amount of defects not present in many of the works in the state of the art. The more important ones are the presence of garbage and detritus as well as the fact that many of the organisms are overlapping each other. These features of the dataset makes both segmentation and classification much more difficult.

Overall the picture quality of the images is really good with a high resolution and clear colours, but they have some caveats. The illumination, due to the lamp position in the microscope, is not even. This will be further discussed in chapter 4 as it will have a big impact on foreground-background segmentation techniques. The trade-off that comes with the higher image quality is the time that needs to dedicated to capturing this stills. The robotic approaches described before are able to take images of plankton directly into the water reducing the time need to gather and process samples. This comes at the cost of image quality, price, size and the water samples themselves as the images are capture from in-motion water that is not saved. FlowCam requires the gathering the samples that would later be introduced in the machine. This approach creates higher quality images when compared to the robotic approaches as they have more resolution and full colour. Finally the microscope approach is the most time consuming but has the advantage in image quality. This is due to the total control of the process by the expert. It is also the cheapest option by a wide margin since the prices of the special hardware can reach hundreds of thousands of dollars. The higher image quality allows the proposed system to extract clearer features, and together with price and easiness of deployability is one of the main factors when motivating this work.

Intraclass morphological differences

Phytoplankton, due to their nature, are inheritable flexible. They exhibit different shapes depending on the axis from which the organisms are observed, which also further digs into this problem. Thus, two specimens of the same species may look completely different, which is an added difficulty for the classification. For example, in figure 2.4 two specimens from the same species, *Woronichinia naegeliana*, are shown. First and foremost colour of the specimen in the 2.4a image is clearly more green while that in 2.4b is quite darker. The shape is also very dissimilar. It is also relevant to note that 2.5a has visible holes through which the background can be appreciated, while 2.5b is more dense in appearance.

(a)

(b)

Figure 2.4: Example of two dissimilar specimens of Woronichinia naegeliana

(a)

(b)

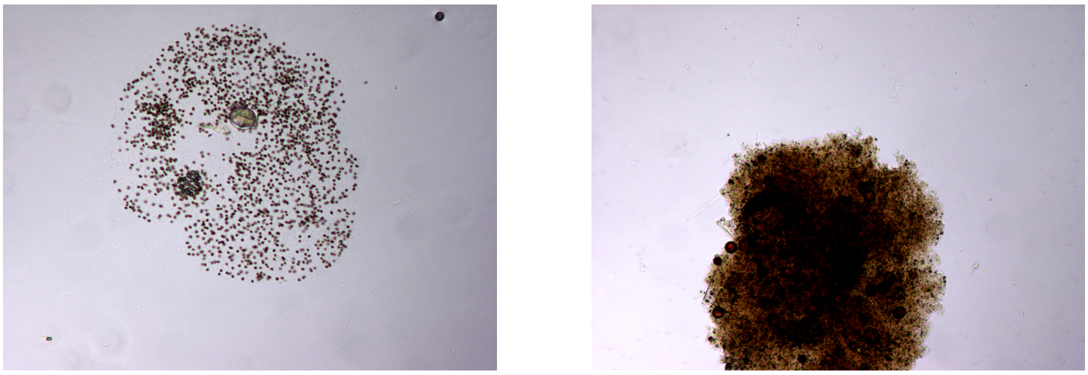
Figure 2.5: Two similar organisms belonging to different species

Interclass morphological similarities

While some specimens from the same species might be really different, it is also common for specimens from different species to be really similar. This also poses difficulties for the classification. Figure 2.5 shows specimens of *W. naegeliana* and *Microcystis flos-aquae*, that are similar to each other. Both individuals share common shape and colours. In terms of texture, *M. flos-aquae* has a pattern that is shared by many *W. naegeliana*, consisting on several smaller dots around the organism. This feature is a little blurry in figure 2.5a but can be better appreciated in the example of figure 2.4.

Specimen overlapping and superposition

In the images for this project it is quite common that some specimens are partially overlapping with others, or even that some specimens appear completely within some bigger organism. This can be clearly seen in the example of figure 2.6a. It is observed that several organisms are completely overlapping a bigger one, which makes the individual specimens nearly impossible to separate. In addition, the resulting image with the overlapping specimens will be difficult to classify into either of the species due to the overlapping features of both classes.



(a) Several organisms like *Dinobryon Sp.* inside a *Microcystis aeruginosa*

(b) Great garbage accumulation over different organisms

Figure 2.6: Example of difficulties in terms of segmenting specimens on the dataset

High amount of garbage and imaging artefacts

The set of images are from real world water samples. Thus they contain all kinds of spurious elements. Notably, the collected images have high amounts of trash in them. Some of them are even completely filled with it, like in the example of figure 2.6b where the individual organism segmentation and detection is pretty much impossible. This general dirtiness of the images difficults the process of the candidate detection by introducing loads of detections that are not phytoplankton. While it makes sense for the system to pick up these candidates due to their organism-like is appearance, they fill up the data with noise. This complicates classification since the garbage can come in any shape, colour, etc., that can sometimes very closely resemble the real specimens.

There are different groups in which the spurious elements that are present in the dataset could be broken down. Each category is represented in the figure 2.7. In that figure it is easy to see that both organic garbage 2.7c and zooplankton 2.7d are really similar to phytoplankton. Their similitude encompasses many features, from colour to shape, size or texture. Those two groups are the most difficult to separate from phytoplanktonic organisms.

Bubbles present in 2.7b can also be similar to some species of phytoplankton, as they often share a rounded shape an can appear to contain some material other than air. Shadows, like shown in figure 2.7a can also be similar to some specimens due to their shape and dark colours which can be alike a *Woronichinia naegeliana*.

Inorganic garbage includes, but it is not limited to, plastics and cloth fibres like the one in figure 2.7e. While their features are more distinct to microorganisms that those described above it is clear that some of them are still quite similar to phytoplankton. This is specially true for some mineral and rock particles that may appear green and have a circular shape, just like many specimens. The translucent colours are shared with some of the imaging defects.

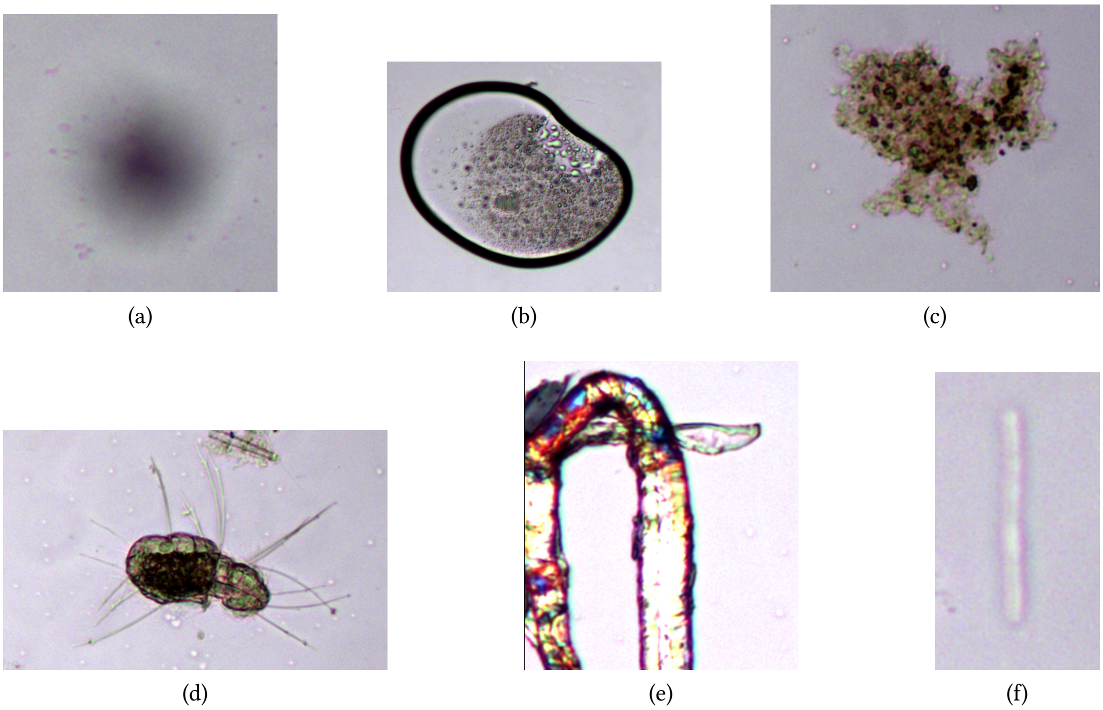


Figure 2.7: Different types of artefacts and defects present in the images

While the latter are less similar to organisms both these groups could be confused with the transparent membranes that some species of phytoplankton like *Dinobryon Sp.* possess. This is accentuated if some of this garbage types are combined, like for example, if a sand particle appears to be inside a transparent plastic it would be very difficult to differentiate from a normal phytoplankton specimen.

Imaging artefacts like the one in figure 2.7f can share some features with some phytoplankton species and their shape can, sometimes, be deceptive.

2.3.3 Relation to previous work

In general, while there are a lot of different works focusing on in the plankton image acquisition and analysis, the main focus of those is still mainly saltwater plankton, which is divided between zooplankton and phytoplankton research. Freshwater remains relatively unpopulated by works and there is no one with the same intentions and pipeline as the work herein described, making this work innovative. The fact that it aims to complete a the analysis from beginning to end without special hardware and on freshwater represents serious originality when compared to other state of the art works.

The main difference between this project and many others [25, 37, 46] is the fact that this is a full, complete pipeline from segmentation to classification. Most of the works cited before are only focused on classification, and a lot of them use a premade dataset that consists on images from one of the automatic capturing devices described in the section 2.2. These sets of images are ideal, selected ones, lacking the complexity characteristic of the images used in this project as they just show one individual and are free from garbage and other common defects. The lack of real world features limits the difficulty but also limits the real-world use of the developed technologies. The images used in this work contain a set of complex features described in section 2.3.2 making the system a realistic approach. This features add up to the already complicated problem that the intraspecies variation but interspecies similarity pose in phytoplankton classification.

Chapter 3

Planning

In this section of the dissertation the planning set for the project will be explained together with a breakdown of human and material resource costs.

3.1 Development Methodology

As this work is of research type it is known that it will suffer from continuous revisions and modifications depending on the obtained results, which may change the direction of the work. Knowing this, the model that is more fitting is the incremental development model [64], as it adjusts well to this type of conditions allowing for evolution and change along the project development. The incremental development model allows to create a working product at the end of each iteration. Which, in turn, allows for a project that is developed step by step to gain constant feedback. The process starts with an analysis and ends with testing of the developed product. The design and codification steps are in between these two, as it can be seen in figure 3.1 [65].

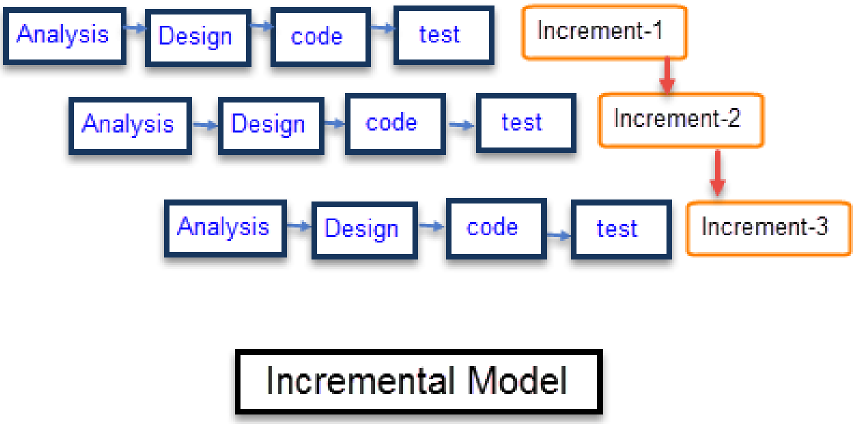


Figure 3.1: Incremental development model

3.2 Tasks

At the beginning of the planning stage it is necessary to establish a set of tasks or activities that will be carried out during the development of the project. They can be broken down as:

1. **Study of the domain:** At the beginning of the project it is necessary to study the context and domain surrounding the project to better understand its importance.
    - **Study of the biological scope:** An important part of the study of the domain is to collect information related to the biological part of this work: what is plankton and phytoplankton, why it is so important for the environment and human health, how can it become toxic, and how dangerous it is, etc.
    - **Bibliographical study:** The other part of this task will be focused on computer science as it is needed to know what research and systems have already been made and what level of progress there is on the field.
2. **Segmentation and detection:**
    - **Creation of a dataset:** Before any relevant work can be done, the available data needs to be sorted and organised. It is fundamental to do so as as the data will be the base for the sample analysis. The dataset consists in microscopy images from real water samples.
    - **Creation of a segmentation methodology:** Development of a computer vision methodology that allows to segment the images into background and foreground and detect phytoplankton candidate specimens.
    - **Experimentation and evaluation:** Extensive testing and revisions on the methodology until satisfactory results are gathered.
3. **Specimen classification:** After the specimens are segmented, it is necessary to classify them. This requires creating a new dataset from those specimens and tagging them.
    - **Creation of a dataset:** From the results of the previous step a new dataset is created and tagged appropriately.
    - **Separation between phytoplankton and others:** This step intends to recognise which of the candidates are phytoplankton organisms and which are other organisms or debris.
    - **Experimentation and evaluation:** Following the previous subtask tests and experiments are conducted to test and improve the methodology.
    - **Species classification:** The focus is to detect dangerous cyanobacteria in water to evaluate their amount and thus assess the potability of the water. In this subtask a classification is preformed to differentiate between harmless and harmful phytoplankton.
    - **Experimentation and evaluation:** Following the previous subtask tests and experiments are conducted to test and improve the methodology.
4. **Elaboration of the dissertation:** Once the work is finished it is time to write a dissertation explaining the details of the work.

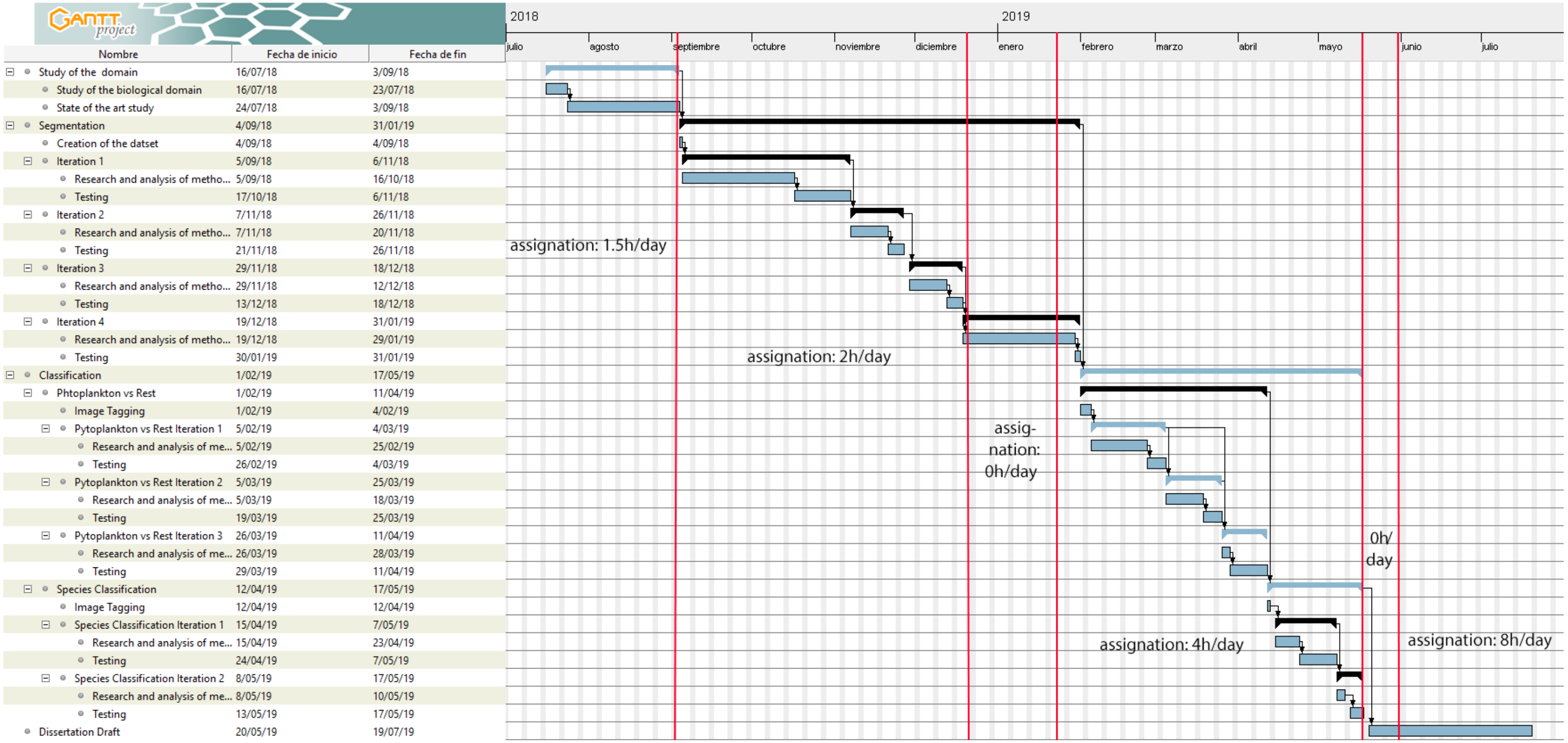
3.3 Project schedule

Figure 3.2 shows a Gantt diagram for the project that establishes a baseline schedule for this work, with the purpose of helping with the timing of the project. The planning includes several tasks divided in iterations following the criteria of the development process. This allows to mitigate the effects of the unpredictability of the results and for a correct result evaluation and correction of the developed modules. Following this planning, the work started on July 2018 and finished on July 2019. Research projects expose them to extreme deviations from the planning as experimental results or new information is gathered. The project is also carried out together with the normal subjects of the fourth year of the degree, which aggravates the planning problems, as their dense contents can normally occupy the student's full time.

This situation motivated that the project is divided into several periods with varying work hours per day. The first part, which mainly consists on investigating the domain is developed in the summer of 2018. A workload of one hour and a half a day is assumed. Then, until February 2019 the work is performed during a quarter of workday. This explains the longer duration of tasks as the resource assignation was less than desirable if compared to a normal 8 hour workday. For the second half of the year it is supposed that the work is carried out in half a workday. Finally after May exams, it is assumed that the work was carried out in a full-time manner. During the exam period, late December-January and May the tasks were completely stopped to focus only on studying for said exams. This is the reason for some tasks to appear disproportionate in figure 3.2. The tasks affected by this stoppage are longer than their counterparts due to the lack of assignation during many days. Specially visible in figure 3.2 is the research and analysis of methodologies on the fourth iteration of the segmentation process. This task had no assignation during the end of December and most part of January resulting in a longer spawn of time. A similar thing can be appreciated in the dissertation draft tasks, that due to the lack of assignation ends up spawning nearly two weeks longer, a shorter time than in the first quarter due to the lesser amount of subjects in the second one.

3.4 Project execution

After the execution of the project, the main deviations were found in the second half of the schedule. While in the first half everything went, more or less, according to the schedule the second half proved to be more difficult to plan. The first tasks of the segmentation process came along nicely with the computer vision subject which helped reduce the research time a little. This enabled the experimentation with more preliminary solutions that did not work, instead of simply overlooking them. The initial planning of classification step was on the other hand, underestimated. For example, some unexpected delays were experienced during the



Nombre	Fecha de inicio	Fecha de fin
⊟ Study of the domain	16/07/18	3/09/18
Study of the biological domain	16/07/18	23/07/18
State of the art study	24/07/18	3/09/18
⊟ Segmentation	4/09/18	31/01/19
Creation of the datset	4/09/18	4/09/18
⊟ Iteration 1	5/09/18	6/11/18
Research and analysis of metho...	5/09/18	16/10/18
Testing	17/10/18	6/11/18
⊟ Iteration 2	7/11/18	26/11/18
Research and analysis of metho...	7/11/18	20/11/18
Testing	21/11/18	26/11/18
⊟ Iteration 3	29/11/18	18/12/18
Research and analysis of metho...	29/11/18	12/12/18
Testing	13/12/18	18/12/18
⊟ Iteration 4	19/12/18	31/01/19
Research and analysis of metho...	19/12/18	29/01/19
Testing	30/01/19	31/01/19
⊟ Classification	1/02/19	17/05/19
⊟ Phtoplankton vs Rest	1/02/19	11/04/19
Image Tagging	1/02/19	4/02/19
⊟ Pytoplankton vs Rest Iteration 1	5/02/19	4/03/19
Research and analysis of me...	5/02/19	25/02/19
Testing	26/02/19	4/03/19
⊟ Pytoplankton vs Rest Iteration 2	5/03/19	25/03/19
Research and analysis of me...	5/03/19	18/03/19
Testing	19/03/19	25/03/19
⊟ Pytoplankton vs Rest Iteration 3	26/03/19	11/04/19
Research and analysis of me...	26/03/19	28/03/19
Testing	29/03/19	11/04/19
⊟ Species Classification	12/04/19	17/05/19
Image Tagging	12/04/19	12/04/19
⊟ Species Classification Iteration 1	15/04/19	7/05/19
Research and analysis of me...	15/04/19	23/04/19
Testing	24/04/19	7/05/19
⊟ Species Classification Iteration 2	8/05/19	17/05/19
Research and analysis of me...	8/05/19	10/05/19
Testing	13/05/19	17/05/19
Dissertation Draft	20/05/19	19/07/19

Figure 3.2: Gantt diagram with project baseline

creation of the classification dataset due to the need for coordination with the expert biologist. Nevertheless, the most important schedule bias was the time taken to run the experiments. Due to the computational complexity of the techniques used and the big amount of tests, VARPA servers were used to speed up the computational process increasing the amount of parallelisation. This bias produced a delay in the drafting of the dissertation, which started without having the testing part of the last step completely finished. This, however, did not impact the development of the project in any way and the parallelisation of both tasks served to save some time.

3.5 Resources

The resources that were used in this project are broke down below as well as classified into two categories: human resources and material resources.

3.5.1 Human resources

The different roles of the people involved in this project:

- **Project manager:** The leader of all the people involved, responsible for the planning and execution of the project. This role is shared between the directors and the student.
- **Developer:** Also known as analyst or programmer. Is in charge of analysing the main objectives of the project and creating the system itself. This role is fulfilled by the student.
- **Expert biologist:** Responsible to solve doubts that a team member may have regarding the topic of the work. This role is performed by the collaborator biologist, Rafael.

3.5.2 Material resources

The material resources can be divided into two main sections: hardware and software.

Hardware:	Laptop: MacBook pro retina mid 2015	Desktop
Processor	Intel Core i5-5257U	Intel i7 4790K
RAM	8 GB DDR3	16 GB DDR3
GPU	Intel Iris Graphics 6100	Nvidia GeForce 780Ti

A couple or servers belonging to the VARPA group were used to speed up the computation of the tests. These servers were equipped with an Intel Xeon CPU E5-2650 v4 @ 2.20GHz processor, 128GB of RAM and two Tesla K80 GPUs

Software:

- **Laptop OS:** macOS 10.14 Mojave
- **Desktop OS:** Windows 10 Home x64
- **Server OS:** Debian
- **Python 3.7**

- **OpenCV:** Library focused on computer vision. Based on C++ API that provides faster speeds than python-based implementations.
- **Sklearn:** Library focused on machine learning based on other python libraries like NumPy and SciPy.
- **Scikit Image:** Library for image processing and computer vision.
- **SciPy**
- **NumPy**
- **Matplotlib**
- **Joblib**
- **Gantt Project**

3.6 Cost estimation

In this section the cost of the whole project is estimated. Once again the resources are divided into human resources and material resources.

3.6.1 Estimated costs of the human resources

To calculate the total costs of the human resources, all the people involved in the project must be taken into account, what their normal salary would be, and their hours worked in the project. With the directors, a weekly meeting was done with an estimated run time of 2 hours. This results in a total of 72 hours, which should be multiplied by two people. In the case of the biologist, a bi-weekly meeting is scheduled also with a duration of 2 hours. This meeting is simultaneous with the directors one. For the analyst a normal 5 day week is supposed with the hourly assignation described before, a total of approximately 1200 hours is reached. Assuming then a hourly salary of 25€/h for the analyst and 35€/h for the project managers as well as the expert, the total cost can be viewed in the table 3.1.

3.6.2 Estimated costs of the material resources

Like previously, material resources are broken down into two categories:

- **Hardware:** These costs could be dismissed as all the systems were owned before the start of the project.
- **Software:** This costs could also be dismissed as python and its libraries are free and open source. The same can be applied to GanttProject, the program used for the planning.

	People	Wage	Hours/person	Cost
Project Manager	2	35€/h	72	5040€
Domain Expert	1	35€/h	36	1260€
Analyst	1	25€/h	1200	30000€
			Total	36300€

Table 3.1: Human resources cost estimation

Chapter 4

Image segmentation

This section details segmentation the methodology and the results obtained from the evaluation of this stage. The method is divided into two stages. The first consists on separating the candidate specimens from the background. The second attempts to merge specimens into colonies, which is a common way for phytoplankton to appear. After the detailed description of the methods, the experimental results are shown and discussed.

4.1 Overview of the method:

This chapter is focused on detecting the specimens of phytoplankton from the digital microscope images described in the section 2.3.2. This is a complex task that can be broken down in several subsequent stages. The first one is the separation of background and foreground of the image. This allows to detect all the candidate specimens that could be phytoplankton differentiating them from the rest of the image that holds no interest. Following this task the candidates that do not meet certain criteria, like size, are removed as they are not phytoplankton. Phytoplankton often appears in colonies and this is accounted in the next step that fuses similar specimens into colonies, even if visually they hold no connection.

4.2 Dataset

The dataset used for the evaluation of the segmentation and the detection stage, described in this chapter, is composed of images from a digital microscope. Each one of these images, already described in section 2.3.2 feature many phytoplankton specimens as well as a varied amount of garbage. The only change that this images suffered from the direct output of the microscope camera's is a re-sizing from an original size of 2080×1540 pixels to a a more manageable 1000×740 pixels. This allowed the testing process to become quicker without sacrificing image quality. The set of images is composed of a total of 211 images. They are separated into test and training sets. The training set consists of 50 images chosen at random while the test set is composed of the remainder of the images. The different methodologies will be developed by adjusting the parameters to the training set. This way the parameter tuning is not influenced by the test set and thus the result obtained from it are free of any overfitting. The methods are tested over the test set where the relevant metrics for the evaluation are computed.

It is important to note that, every image has a varying amount of specimens. While some may only have a couple or a few other can present up to dozens. This is innovative as it is rarely done but allows to reduce even further the work of the specialists as they have to take less pictures since the magnification is smaller.

The ground truth for this step is composed of the specimens that should be detected. This encompasses every element present in the image that can be considered foreground, that is, everything that could possibly be a phytoplankton specimen. This, however, does not include many elements that may appear to be foreground but are not, like imaging defects and shadows, which can be quite similar to plankton. This is done with the intention of detecting the highest amount possible of phytoplankton specimens as the precision in the final measurement of cyanobacterium is key. This means that detecting every phytoplankton instance and not *losing* any throughout the different steps is really important.

The main objective for this step is, then, to detect every phytoplankton specimen missing the least ones, even if the results include some noise. The noise is inevitable due to similarities between phytoplankton and some of the other elements present in the dataset such as zooplankton or garbage. This noise will be cleaned up going forward, a process which is detailed in chapter 5.

4.3 Foreground-Background separation

This task intends to separate the background of the image, that holds no interest, from the foreground. The foreground can include a big variety of objects since the collection of images is real world data. The foreground includes phytoplankton and other things like zooplankton, garbage organic or not, sand particles, etc. Due to the sampling and imaging process, the images also have other kind of imperfections, such as as shadows, uneven illumination, etc. The simpler one to see is bubbles, but given their shape and colour can also be easily be confused with microorganisms.

The chosen and tested methodologies need to accommodate to these dataset features. The system needs to be as robust as possible since the presence of these spurious elements in the images is mostly random. These imperfections promoted that many techniques and configurations needed to be tested.

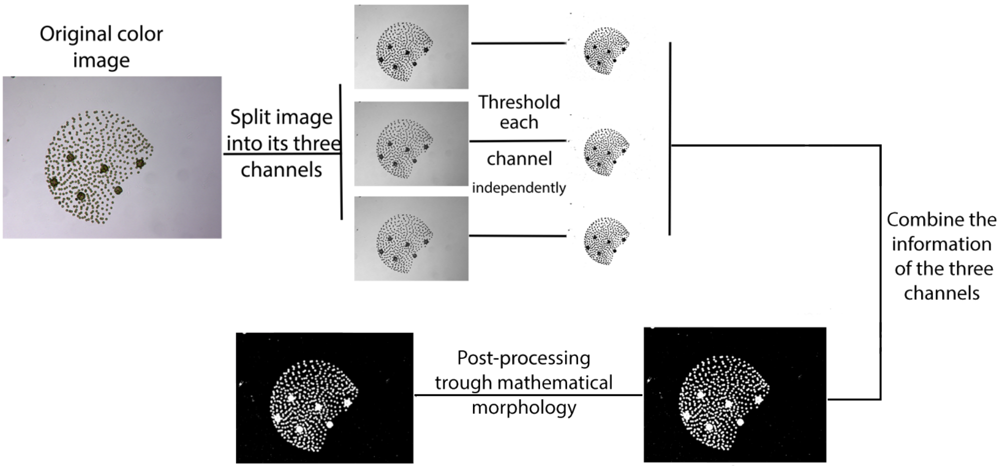

Original color image
Split image into its three channels
Threshold each channel independently
Combine the information of the three channels
Post-processing trough mathematical morphology

Figure 4.1: Overall vision of the process of Foreground-Background separation

A general overview of this section can be seen in the figure 4.1 that shows the steps that will be taken to solve this part.

4.3.1 Global thresholding

The first tested methodologies are the global thresholding methods. Thresholding is a segmentation process in which an image, usually in grayscale, is binarised. Grayscale images represent each pixel with a intensity, 0 for black or the absence of colour and usually 255 for white, the highest possible amount of colour. A binary image just has two values, 0 for black and 1 for white, another way to put it is that a binary image has two classes, background and foreground.

Fixed threshold method

The first tested method is the use of a fixed global threshold. All the pixels value under the threshold value are assigned to background, and to foreground otherwise. This results on a binary image that depends on the chosen threshold value. Figure 4.2 shows a global thresholding example with two different fixed thresholding values. Depending on the selected threshold value the results change drastically, which makes the tuning of this algorithm quite hard for a varied set of images. This is the reason it was abandoned quickly, as either it selected too much noise or it missed candidate parts or even entire organisms.

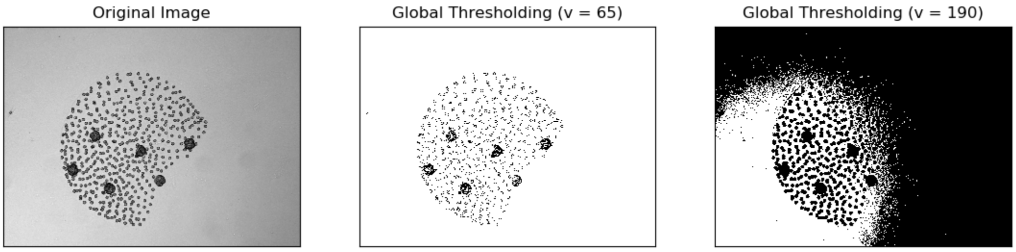


Figure 4.2: Fixed global threshold examples with different values

Otsu's Method

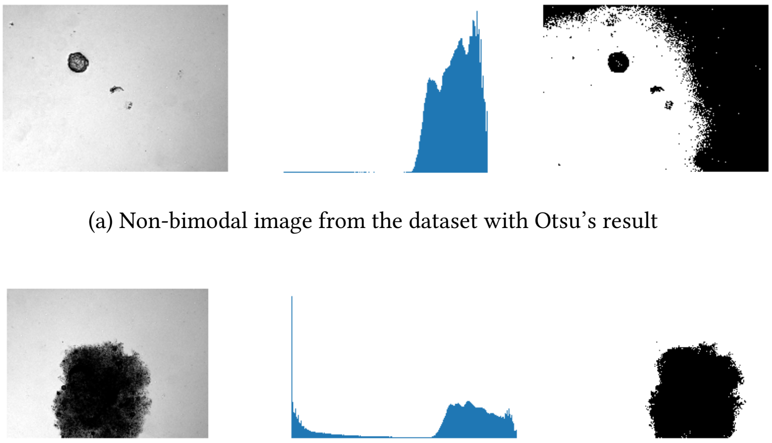
Otsu's method is a more advanced global thresholding method, that allows to automatically select a threshold value based on the image histogram contents. It assumes two classes, which would be background and foreground pixels, and separates them using the threshold value that minimises the intraclass variance. More information about this method can be found in appendix A.1. The main issue is that this algorithm only works well if the image has a clear bimodal histogram. Most of the images from the dataset do not have this feature and thus Otsu will not perform to its optimum. For an image in the dataset to have such a histogram the amount of background and target foreground should be roughly similar. However, in the most of images the histogram only has one peak due the high amount of background. This means that depending on the number of microorganisms in the images this technique would perform well or not. Th density of microorganisms in the images can not be controlled in advance, in real world scenarios. In the figure 4.3 an example of both a bimodal and a non-bimodal image can be seen, both from the dataset. In the non-bimodal image, figure 4.3a, the segmentation is rather lacklustre picking up a lot of background as organisms. On the other hand, the bimodal image in figure 4.3b shows a near optimal segmentation that picks up all the garbage and specimens inside of it as foreground.

4.3.2 Global and local contrast enhancement

Most of the images in the dataset present uneven illumination. As it can be observed in the example of 4.2 a global threshold can not adjust effectively to the varying illumination conditions across the image. The first approximations tried to correct the illumination defects using several types of image equalisation like Contrast Limited Adaptive Histogram Equalisation (CLAHE) [66]. These techniques are explained in greater detail in appendix A.2

Adaptive histogram equalisation should, in theory, improve the performance of global thresholds due to the local operation of the algorithm in cases of uneven illumination.

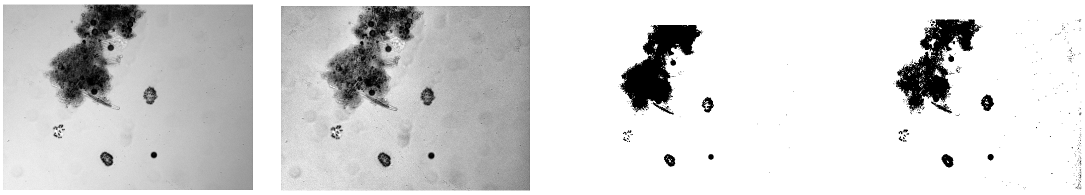
While CLAHE did improve the performance of some of the images where Otsu was useful,



(a) Non-bimodal image from the dataset with Otsu's result

(b) Bimodal image from the dataset with Otsu's result

Figure 4.3: Comparison of Otsu's method results in images with different shaped histograms



(a) Original image

(b) CLAHE image

(c) Segmentation of the original image

(d) Segmentation of the CLAHE image

Figure 4.4: Demonstration of CLAHE and segmentation over a dataset image

however, it did not manage to improve the vast majority. An example of this can be seen in the figure 4.4. The CLAHE image is clearly different from the original one, 4.4b and 4.4a respectively, but neither of the results manages a optimal segmentation. The segmentation derived from the original image, in figure 4.4c, misses parts of the blob of trash that are, coincidentally, specimens. The thresholding derived from the CLAHE image, in figure 4.4d, misses more of the contents of the garbage blob while highlighting some noise in the right part of the image.

While the adaptive histogram equalisation is able to correct the uneven illumination so that an appropriate global threshold could be selected, the selection of the appropriate threshold in terms of the histogram balance is not plausible due to the variation in the image contents (i.e. the number of microorganisms varies from one image to another). It was concluded then, that even with CLAHE, Otsu's method was not suited for this set of images. Even if it proved to correctly segment some of them its performance could not be improved with CLAHE due to the varying contents of the images.

4.3.3 Adaptive Thresholding

An alternative to using a global threshold is the use of adaptive thresholds that depend on the local image contents. Local methods should, in theory, solve some of the problems that global methods exhibited as they check the image contents locally. That is, to decide the threshold value, this methodologies do not use the whole image as data, rather they use a small patch or window and repeat this process for every patch in the image. A more detailed explanation of this method is located in appendix A.3. These methods are also able to correct and adapt to uneven illumination so they make CLAHE unnecessary.

This local thresholding methods are called adaptive. The main differentiation between them is the metric they use to find the threshold value in the window of the image. The best performing in this step was the Gaussian adaptive threshold that makes use of a Gaussian function to weight the values of the pixels. The other version tested, based on the mean as metric, showed worse performance as it picked up more noise.

4.3.4 Use of colour information

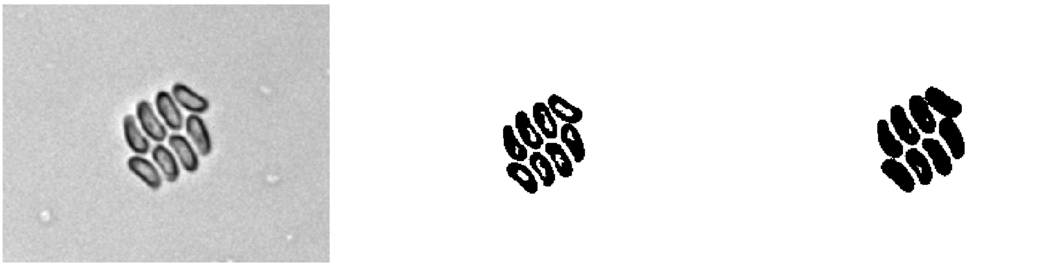
The initial thresholding tests were performed on grayscale images only. The addition of colour information could enable the use of more information and, possibly, better results. Several different approaches, as described below, were tested. A more detailed explanation on the RGB colour model, basis of the colour information, can be found in appendix A.4

RGB green channel

The green channel of the images contains relevant information for phytoplankton. By thresholding on this channel only the intention is detect the green appearance that characterises the phytoplankton organisms. The idea is to try to simulate the methods based on chlorophyll fluorescence described in the state of the art, without actually having any kind of special device. Thresholding only over the green channel, however, reported no benefit, partially because many relevant organisms are quite brown or even black, which differentiates them quite a lot from the green filled ones. Some of the green specimens also showed problems when their hue tends to lighter colours that are a mix between channels, like yellow. An example of this phenomenon can be seen in figure 4.5. The green specimen in figure is segmented badly due to the more faint colours in its inside.

Combination of all RGB channels

In order to integrate other colour information (apart from green), a combined thresholding over all the RGB channels is proposed. This consists on getting the original colour image,



(a) Green specimen green channel (b) Green channel segmentation (c) Grayscale segmentation

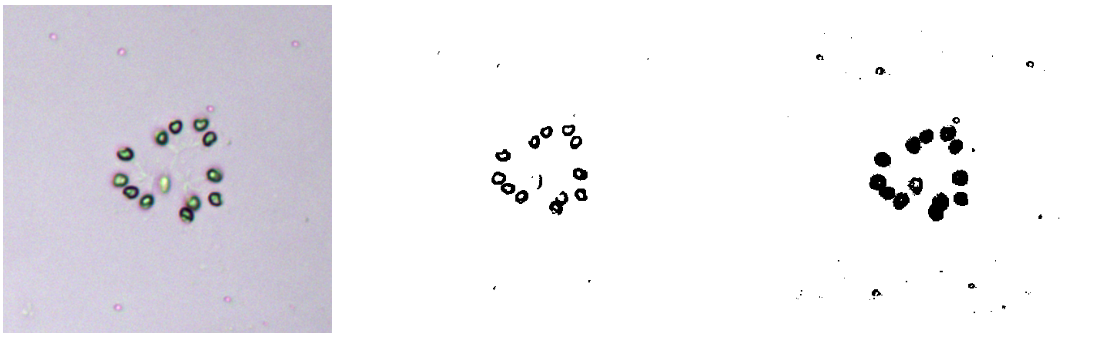
Figure 4.5: A mostly green specimen under different segmentations

splitting it into the three RGB channels and threshold each one separately. The channel-wise foreground segmentation results are then combined through union (OR operation). This allows to integrate the information of any colour in the image. For example in organisms where their colour is dark the value of each channel is going to be low for all three as dark colours tend to 0, but the value for each channel will vary depending on the tone. In the case that an organism is predominantly green the values of the green channel would be higher than those of the other two channels, the same can be applied for some organisms which have a red hue. By using each channel more information is used when compared to the grayscale image, which as explained before, should allow for better specimen detection.

In the figure 4.6 a comparison between methods can be seen. In this particular case, the specimen is mostly green with lighter tones tending to yellow or white inside each separate blob. This proved to be difficult for green channel only as stated in the previous section 4.3.4. However, using all three RGB channels, like shown in figure 4.6c, the result is much improved from the grayscale one, in figure 4.6b. In the grayscale threshold the parts that corresponds to the lighter tones are lost, which is specially apparent in the blob nearing the centre of the specimen. While this issue is still somewhat present with the combination of RGB channels it is of lesser importance. In this images it can also be appreciated that, although RGB is better than just grayscale it picks up more noise around the background. Those marks can be clearly seen in the colour image so it is normal for them to be picked up. This noise can be filtered later without much hassle.

4.3.5 Post-processing

To solve some of the caveats presented in the previous step a post-processing is applied to the results of said step. This post-processing is based on mathematical morphology, which is further explained in appendix A.5. By applying a closing operation over the binary image resulting from the RGB threshold the results can be cleaned up. In order not to modify or alter them significantly a small kernel is chosen. The objective of this step is to fill in the gaps



(a) Original specimen

(b) Grayscale segmentation

(c) Combination of RGB channels segmentation

Figure 4.6: Comparison of grayscale and RGB channels segmentation

(a) Combination of RGB channels segmentation

(b) After post-processing

Figure 4.7: Before and after of post-processing

present in many detections while avoiding any fusion of cells that might be a byproduct. The cell merging will be done later and doing so by morphology would be an error since it would only take into account distance. An example of this process can be seen in figure 4.7. While not all the holes in the specimen are filled, the post-processing fixes some of them which allows this segmentation to improve towards creating a mask for each image. The holes that are present in the previous results could result in uneven masks that randomly misses part of the organism. While this problem is difficult to fix, this post-processing allows to mitigate it without interfering or worsening any other part of the segmentation.

4.3.6 Review of the foreground-background segmentation method

The objective of the subsystem herein described is to distinguish between foreground and background. The foreground is identified as any object that is in the image, among which the

(a) Original image

(b) Result after foreground-background segmentation

Figure 4.8: Results of the foreground-background segmentation compared to the original image

phytoplankton should be present. The output of this method is a binary mask identifying the image positions that contain foreground.

Many pipelines and methods were tested, but in the end the best results came from the combination of adaptive thresholding over each of the three RGB channels. The detections in each channel are combined via a union to gain every bit of information from every channel.

A general scheme of this step can be seen in figure 4.1 where each step is exemplified. A clearer example of the results from this section, that is, up until this point can be seen in in figure 4.8.

4.4 Initial candidate detection

This step continues where the previous left off. A general overview of the process can be seen in the figure 4.9 that shows the steps that will be taken in this section. The methods in this step aim at refining the background-foreground map obtained in the prior step. This map allows to see connected regions which belong to each individual in the image, an example of this can be found in figure 4.10. In this figure each connected region that is derived from the binary map can be seen in a different colour. To obtain the local image and separate each specimen it is then necessary to process this binary map. This processing mainly consists in detecting each blob and filtering the ones that are considered noise. This allows to create a segmentation mask which, in turn, allows to create the local images for each one of the specimens.

4.4.1 Blob detection

In order to detected each *blob*, an algorithm based on the work of Suzuki and Abe [67] is used. Intuitively this algorithms detects the candidates over the thresholded binary image

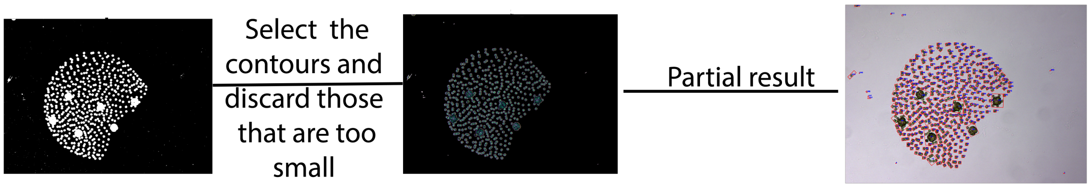


Figure 4.9: Overall vision of the process of candidate detection

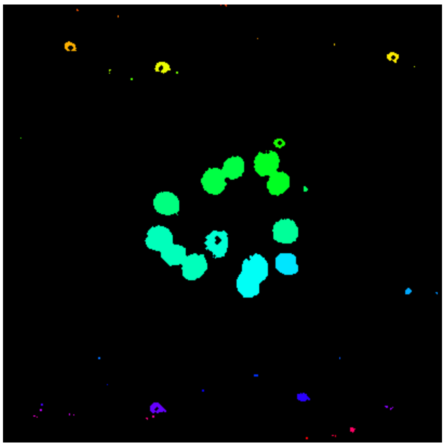
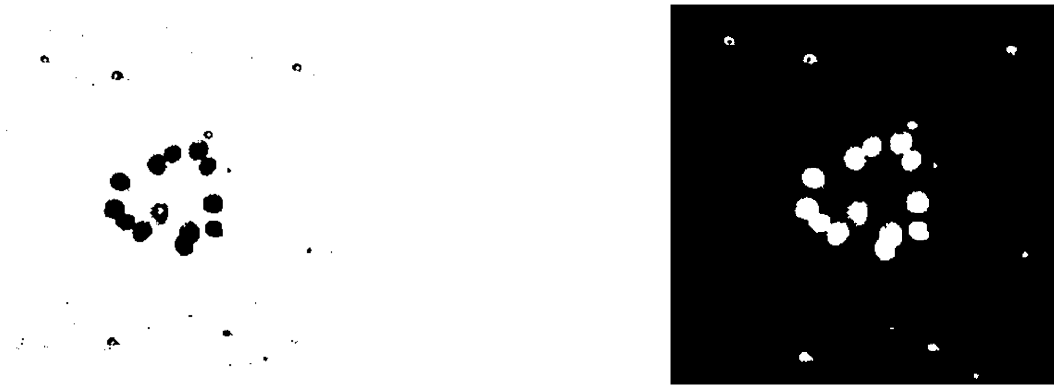


Figure 4.10: Connected components derived from a foreground-background map

separating each one of the masses that were set in the foreground of the image. It works by rasterizing an image and following along its detected borders. The method is also able to differentiate between outer and hole borders of the blobs. By only taking the outer borders into account it is possible to avoid any holes that may appear inside the organisms as a result of the previous segmentation process. A lengthier explanation can be seen in appendix A.6. An example of this process can be seen in the figure 4.11 where each individual blob is detected. As explained before the problem of empty holes in several blobs has been fixed.

4.4.2 Post-processing

Some of the detected candidates consist on small dots and detritus. An expert biologist confirmed that very small specimens i.e. those smaller than 5 µm wide are usually ignored in the analysis. Converting this measure to pixels, by using the known image scale results in 15.5 pixels. Thus it is proposed to use a candidate filtering criterion that removes the specimens not measuring more than 16 pixels (by rounding up). It was decided to apply this criterion over the area since it is a more robust than other metrics. Having an organism with an area of 16 pixels allows it to be of the minimum width (or length) of 1 pixel while still being counted. Using the area, furthermore, allows the smaller circular organisms to be preserved that would



(a) Foreground-background map from previous step

(b) Detected contours

Figure 4.11: Results of the foreground-background segmentation compared to the original image

otherwise not fit the 5 µm criterion. Note that for several organisms the cytoplasm capsule that surround the specimen are not detected due to their transparency, so the segmentation is only able to include their small round nucleus. Finally, using the area means the specimens can adopt any shape while using the same criterion. Using a fixed policy makes the system reproducible and solves the problem that comes from different experts having different criteria, which, in turn, comes from a non-exhaustive process that is generally eyeballed.

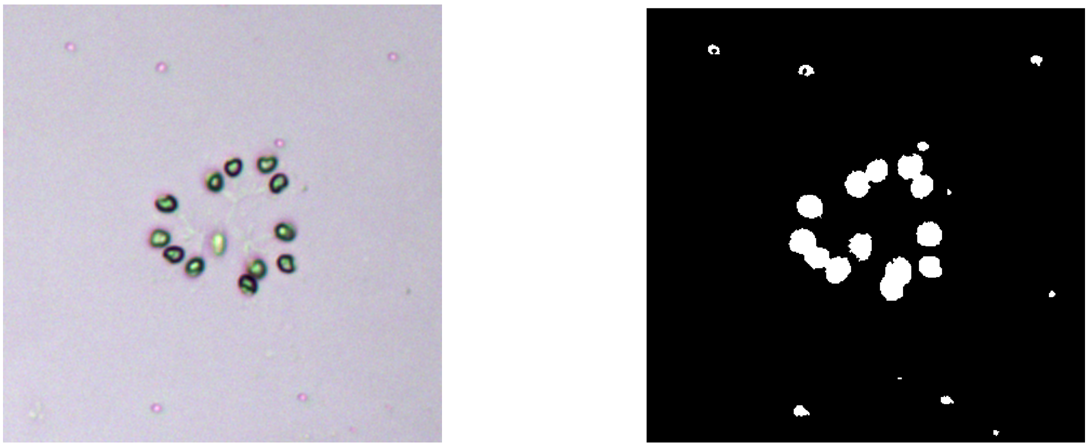
4.4.3 Review of the candidate detection method

This process detects all the contours obtained from the previous steps and filters them to allow only the possible candidates. This process begins with an algorithm that is able to detect all the blobs of the binarised image obtained from the the foreground-background segmentation.

The blobs that are of small size are filtered out following the biologists criteria that specimens too small are not taken into account. This allows to discard all the detected blobs that are not relevant due to their small size. An example of the product of this can be seen in figure 4.12. The specimen in the original image is just one, with minor noise surrounding it. Even if in the output image some of the noise is present, every part of the specimen is fully detected although it is counted separately as the bonds between blobs are too subtle for the method. This is solved in the next step trough merging.

4.5 Merging of colonies

The candidate detection allows to get partial results that are valid for a lot of species with a dense appearance. There are, however, some species that appear in the image in the form of



(a) Original specimen

(b) Output of this step

Figure 4.12: Results of the foreground-background segmentation compared to the original image

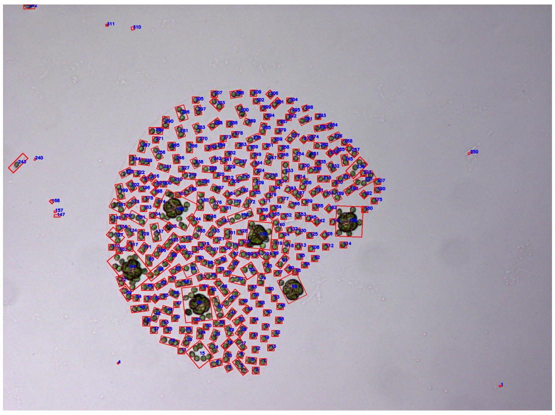


Figure 4.13: Partial results of the methodology previous to colony merging

colonies with sparse appearance. An example of this are specimens of *Volvox aureus*, which is formed by several more or less sparse cells, a phytoplankton organism of considerable size (compared with other species in the dataset). Figure 4.13 shows an example of the results obtained with the proposed candidate detection method on this species. As it can be observed, the detection algorithm separates each of the composing sparse cells into a separate candidate. This is mainly because the inner parts of the organism are not visibly connected, and the only things that lead to the identification of one individual only are the common appearance of the cells, their closeness and the subtle, almost transparent and invisible, membrane that surrounds the whole specimen.

To fuse different detections it is necessary to know which ones are next to each other. Once the neighbouring candidates are established they should be analysed to determine if

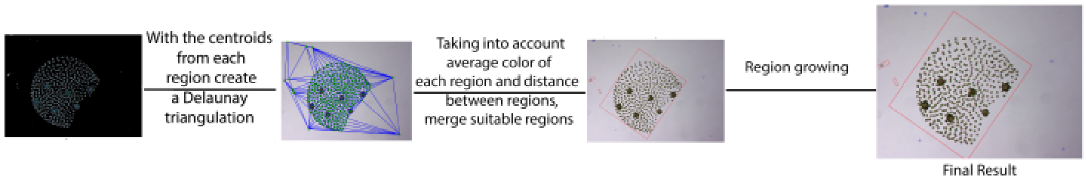

With the centroids from each region create a Delaunay triangulation
Taking into account average color of each region and distance between regions, merge suitable regions
Region growing
Final Result

Figure 4.14: Overall vision of the merging colonies process

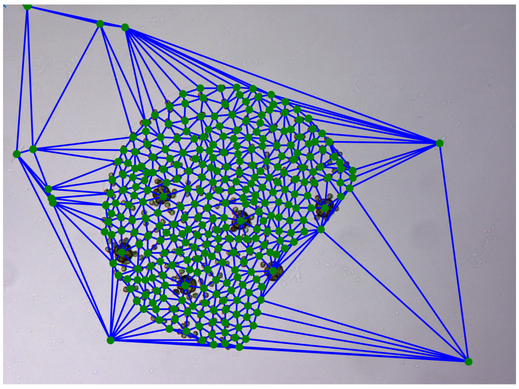


Figure 4.15: Example of a Delaunay triangulation over a image of the dataset

they are similar enough to be fused. The similarity criterion could be based of a various things like colour, shape, size, etc.

An overview of the methods used and the results produced in this section can be seen in the figure 4.14.

4.5.1 Construction of the neighbour graph

The problem of knowing which detections are neighbours of which other ones can be solved using a Delaunay triangulation [68]. The centroids of each one of the candidates is considered as a vertex and a Delaunay Triangulation is performed with them. A lengthier explanation of this method can be found in appendix A.7.

An example of the Delaunay triangulation results over the same *Volvox A.* of figure 4.13 can be seen in figure 4.15. It is observed that every detection is connected to several others forming triangles. This graph can serve as the basis to asses the pairwise similarity of the candidates through the exploration of its edges.

It should be highlighted that to build a Delaunay triangulation, at least four detections are needed so, in order to ensure that the restriction is met, if the analysed image does not have enough detections, this process is simply ignored.

4.5.2 Graph Pruning

Exploring the previously created graph involves taking a centroid or vertex and analysing, through the edges, its neighbours. In this process if both vertexes being compared are similar the edge is kept signalling that they should be fused. If they are too dissimilar and should not be fused the edge is then broken or pruned. It is then important to find a correct metric or combination of them to evaluate the similarity between candidates and whether or not they should be fused. Once every vertex has been visited the different groups that need to be merged should remain connected. These groups are then converted from a combination of smaller blobs into a bigger candidate that encompasses all of them.

To explore every node of the triangulation it is necessary to follow a policy or a set of guidelines. They would allow for a clear, repeatable process. First of all, each point will only be compared with its immediate neighbours. This allows to reduce computations as it would only link two edges that are not neighbours only if they have another vertex acting as a bridge between them. It would also mean that no fusions are carried out if something is physically separating two similar cells or specimens. This policy can be demonstrated or better explained through an intuitive example.*Let there be 3 collinear points named A, B and C respectively. An arbitrary policy determines that AB and BC must be fused. As they share a point, in this case, B AB and BC are subsequently fused again into ABC.* As the example shows it is way quicker to check twice than check A with B and C and then B with C. This has to be multiplied by the great amount of points that are in any given image.

A set of metrics to evaluate similarity between the edges is fundamental for this step. The chosen ones are colour similarity and distance. Colour similarity intends to evaluate how similar the colours of each vertex are to one another. Similar colours between detections that are also close usually indicate that they are from the same species, and thus belong to the same colony. If two detections are similar in colour but are too distant they should not be fused.

The colour metric for each region is obtained with the help of the binary masks created in the previous step. This is done by masking the original image with said mask. By only getting the colours inside of each blob it is avoided that background skews the distribution of colours. Several ways of computing the colour average for each region were subsequently tested. One of them was using the k-means clustering algorithm to get the dominant colour. This method works by passing the information of colour to the k-means algorithm and after it clusters the points into the k groups, the group with the most members will be the dominant colour being the centre its value. The more groups the more detail the colour will have thus reducing the number of members it has. The best performance was, however, obtained from simpler method, a normal arithmetic average. This average was computed on a per RGB channel basis.

The distance metric is computed between the vertexes, that is, between the centroids of each region. It is calculated in pixels.

To fuse the regions their data is checked with their neighbours. If their per channel average colour is within the margin and they are sufficiently close they are fused. It should be noted that both the colour similarity threshold and the distance threshold were adjusted, through experimentation, to obtain the best performance only over the training set.

Once all the vertexes have been explored those that need to be fused are already grouped. It should be noted that, if only the location of the centroid of each region were used to generate a bounding box the result would probably leave part of the region of interest or ROI out. This is due to the very definition of bounding box as it is the minimum area triangle that can enclose a set of points. If the centroids are that set of points it is very likely that important parts are left out. To solve this problem the set of points passed to the merging algorithm is the four corners of the bounding box for each region to be fused. This ensures that no important information is left out of the segmentation. The result of this step can be seen in the figure 4.16a. The bounding boxes in red have been merged with nearby ones while the ones in blue have been left as they were originally. The *Volvox A.* has been completely fused into a big specimen instead of the multiple detections that were seen before.

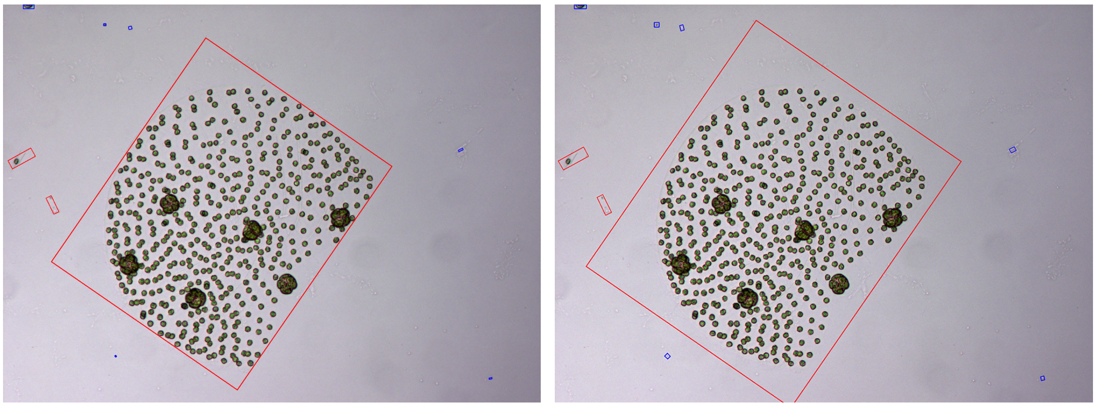
Finally, the generated bounding boxes are grown to include more background information as they might be too tight to the specimen itself. The example of this is shown in 4.16 where it is compared to the not enlarged, original, bounding boxes. The grown part could contain details about the surroundings of the specimen as well as important details of the organisms themselves. That might be specially relevant in some cases where the candidate detections system does not detect the membrane of the organism. This information can be of special importance for identifying the species of the candidates and making the bounding box will, certainly, include it. Each bounding box is grown a 10% in each direction.

4.6 Experimental evaluation

The detection and segmentation methodologies presented in sections 4.3, 4.4 and 4.5 are evaluated using the dataset described in section 4.2. The metrics used in the evaluation are described in section 4.6.1. The tuning process used to adapt the parameters of the methods is detailed in the section 4.6.2. Finally the results are shown and discussed in the sections 4.6.3 and 4.6.4, respectively.

4.6.1 Evaluation metrics

In order to evaluate the performance of the system two different types of metrics will be used. The first ones are focused on evaluating the detection performance. The second set are oriented to analyse the quality of the detected candidates. Both of these metrics are described



(a) Different regions after the merging process, in red regions that have been merged, in blue regions left unmerged

(b) Grown bounding boxes

Figure 4.16: Results of the region merging and bounding box growing

in detail below.

Candidate detection evaluation

Given the ground truth described in section 4.2, the confusion matrix is created between the detected regions and said ground truth. In order to do that it is considered that a positive is achieved when at least 50% of the area of the true specimen is covered by one of the resulting bounding boxes. If the percentage is lower it is said that the particular organism has not been correctly detected and thus it would be accounted as false negative.

In this case true negatives hold no relevance as they would correspond to not detecting the background. As the counting of true positives, false positives and false negatives is done in the form of specimens it makes no sense to count the background as it can not be done using the same metric. The main focus of interest of this work is detecting the foreground, more specifically the phytoplankton organisms so the background holds no further interest.

To evaluate the amount of the detections a confusion matrix is created with special focus on the false negative rate (FNR). A confusion matrix is a concrete way of laying out the data obtained from the experimentation. It presents the predicted data confronted with the real one, allowing for easy comparison between both. From this table several metrics can be derived, like false negative rate that is defined as the amount of false negatives divided by total amount of positives. In this particular case, FNR would be computed dividing the missed elements by the total amount elements that should have been detected. The purpose of using FNR as metric is to limit the misclassified phytoplankton specimens to a minimum as they are the most important point of this work.

Candidate quality evaluation

The different candidates can be analysed by the quality of the bounding box that encloses it. If several bounding boxes are detecting the same organism the organism is identified as oversegmented. For a specimen to be oversegmented more than one bounding box should be trying to enclose it, no matter the size of the boxes nor the area of the organism that they cover. On the other hand it is said that undersegmentation is produced if several specimens are enclosed in just one bounding box. If a resulting candidate box covers more than one organism and each one of those organisms in more than 50% of any of the organisms' size, then that bounding box should have been divided further and undersegmentation is detected. The metrics to evaluate the quality of the candidates are the percentage of specimens that are under and oversegmented over the total of detected organisms.

4.6.2 Parameter tuning

As previously explained the dataset is divided into training and test. The different parameters in the system have been tuned over the training set to minimise the missed phytoplankton organisms that are not detected. That is, the focus of this step is that all the phytoplankton organisms are detected. The reason behind this is that in potability testing is important to count all the malignant organisms to obtain metrics. If some of those organisms are already missed in this step, they can not be taken into consideration in further steps of the analysis. This is done at the expense of obtaining a larger number of false positives. However, these false positives can, and will, be filtered out by the subsequent classification steps in the whole system. It is important to note the fact that during the empirical optimisation of the parameters no test images where used so the estimated evaluation metrics over the test set are not compromised.

The test set was hold out to evaluate the performance of the system.

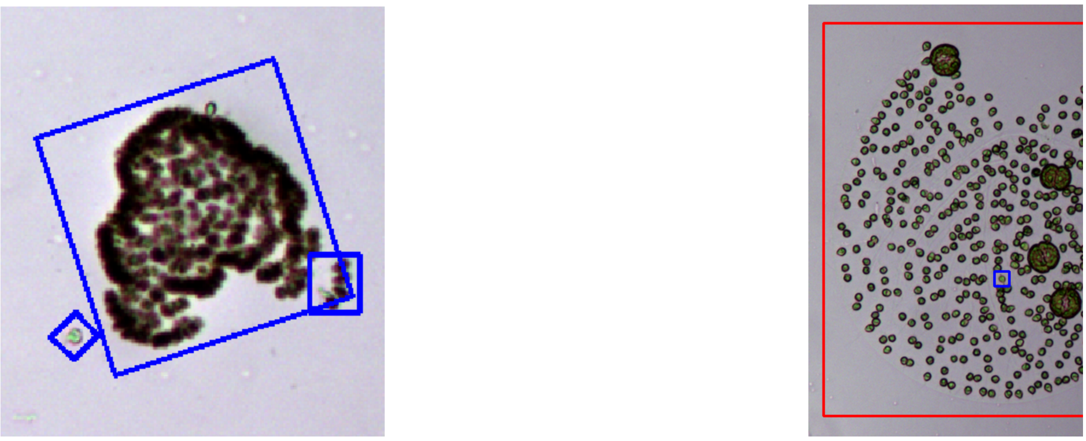
4.6.3 Results

The table 4.1 shows the confusion matrix of the results obtained over the test set comparing the candidate detections with the ground truth. In that table true negatives are ignored, as explained in 4.6.1. The results contained on this confusion matrix yield a 0.256% of FNR.

The metrics for undersegmentation yielded that 12.53% of the candidates were actually undersegmented. On the other hand, only 3.42% of the candidates were found to be oversegmented. Examples of over and undersegmentation can be views in figures 4.17 and 4.18, respectively. In the oversegmentation figure, 4.17, two examples can be seen where the fusion process failed to properly fuse some parts of the organism. In both cases the problem is minimal as most of the organism is still in a bounding box. This is specially important in cases

		Predicted	
		Yes	No
Actual	Yes	1557	4
	No	525	-

Table 4.1: Segmentation results in the test set



(a)

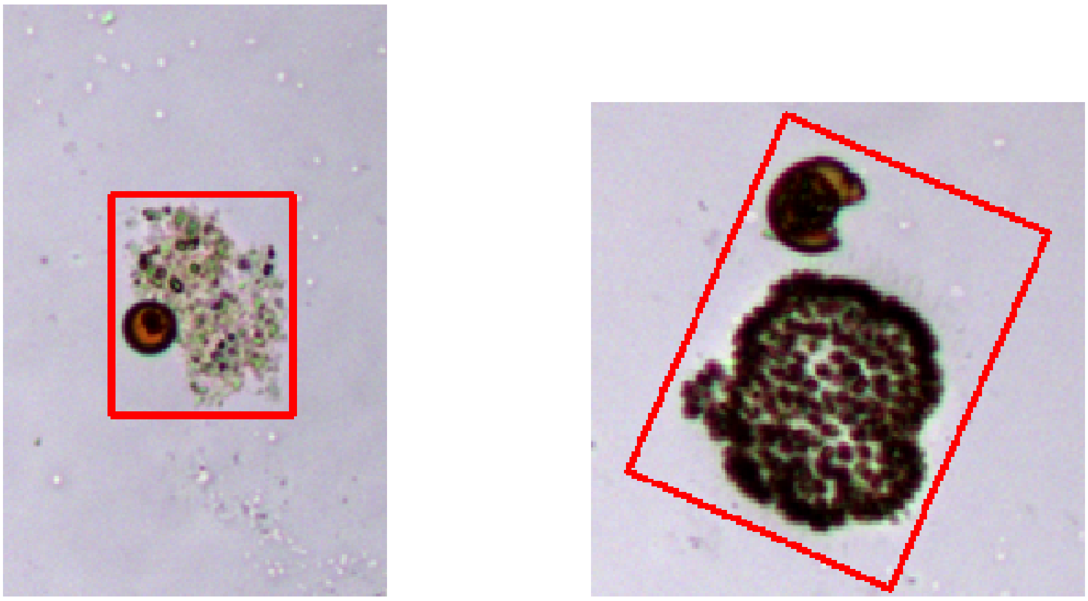
(b)

Figure 4.17: Oversegmentation examples

such as 4.17b where a single part of the organism was not fused due to its different colours. A lot of oversegmentation examples are of this nature and do not affect the detection of the whole specimen as it is complete in a candidate. Undersegmentation is more common in the dataset and its examples can be seen in figure 4.18. In the first example, 4.18a, they closeness of the phytoplankton organism, a *Trachelomonas volvocinopsis*, and the organic garbage is what keeps them fused as they are overlapping. Even if their colours are different. In the second example, 4.18b, the colour is what promotes the fusion between the two specimens, since they are very close but not overlapping.

4.6.4 Discussion

As it can be seen in the presented results, the proposed method manages to not miss only four organisms, which represents a merely 0.256% of false positive rate. The minimisation of this metric was the original objective and all the parameters where optimised to such purpose. It should be emphasised that the detections are according to the metrics described in the section 4.6.1 and if, for example, part of a organism is missed, it can still be counted as a successful detection if more than 50% of its true area is correctly enclosed in the resulting bounding box.



(a)

(b)

Figure 4.18: Undersegmentation examples

It is clear then that with the results previously described, at least the first milestone of the system is complete, missing the least possible organisms.

Looking at under and oversegmentation metrics, the quality of the detections can be analysed. Undersegmentation is an issue that is really hard to mitigate or eliminate due to the images that the dataset contains. It is usual that the objects appear really close together, and even overlapping. This makes the separation of the specimens really difficult, if possible at all. This can easily be appreciated in the example of figure 4.18a. In the example the objects can not be discerned and are detected as a single one due to the overlapping between them. Another reason for the undersegmentation is the excessive merging during the step that aims to group colonies, this can be seen in the example provided in figure 4.18b. This is again due to the small distance between organisms along with the similarities between the fused candidates. This could be improved using more information to decide the graph pruning criteria as, currently, only the distance and colour are taken into account. However, this is a truly difficult task as phytoplankton has a high variability in many variables be it colour, shape or size. Using a more complex fusion algorithm may or may not benefit the system due to this fact.

On the other hand oversegmentation is much less predominant, with a total of 3.42% of the detections. Most of these cases are parts of a colony that should have been fused but they were not. The identifiable reason for this is significant differences in colour with the rest of the colony. An example of this phenomenon can be seen in figure 4.17b where part of the

Volvox Aureus was not fused with the rest of the organism. The example shown in figure 4.17a fails to be fused to the main part of the organism even if their colours are fairly similar. This failure can be blamed on the binary mask that is used to obtain the colours as any errors in such a small individual can lead to parts of the background getting selected and skewing the colour distribution due to the small amount of pixels.

4.7 Summary and conclusions

In this chapter the methodology for the segmentation and detection of specimen candidates in the images was presented and evaluated.

The proposed method starts with an adaptive Gaussian thresholding over the three RGB channels. The segmentation results are them merged together to preserve the most detection. This improves upon using the same threshold method over the grayscale image sensibly. The next process consists on a blob detection algorithm to detect the individuals in the resulting foreground-background segmentation mask, followed by a filter that removes the candidates that are too small to be of relevance. Finally, the candidates that are similar enough in terms of colour and are close to each other are fused together, as they are most likely part of the same colony. The whole process the method follows can be viewed in figure B.6 in the second appendix, which shows every step and the results it produces.

The method showed good performance in the evaluation by only missing 0.256% of the organisms that should have been detected in the whole test set. The system also showed good performance in terms of quality which is evaluated through under and oversegmentation. Oversegmentation was very rare while undersegmentation was more common in the results due to the frequent overlapping between organisms and garbage. Oversegmentation cases, despite less common, were found to be related to the performance of the colony merging algorithm. As future work it is hypothesised that integrating more information into the appearance assessment of the graph pruning method may improve the results.

The main achievement for this chapter is thus the creation of a robust system that can detect and segment freshwater phytoplankton specimen from microscope images that contain several of them, which is something not yet seen in the state of the art.

Chapter 5

Classification

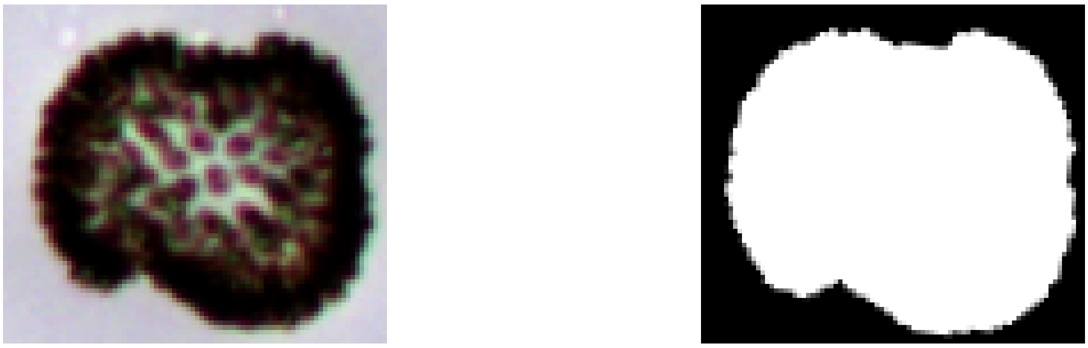
This chapter is focused on detailing the candidate classification methodology and its evaluation. The chapter is split into two parts. The first focuses on the differentiation between phytoplankton organisms from the rest of candidates detected in the previous step. The second part consists on the classification of phytoplankton organisms into different species with the objective of providing a methodology that allows to identify dangerous cyanobacteria.

5.1 Overview and objectives

The methods described in the previous chapter allowed to identify the several specimen candidates that are present in each image. The detection process was optimised to minimise the ratio of false negatives so that all specimens present in the images are retrieved for further processing. This, nevertheless, causes the inclusion of a high number of false positives in the candidate set. Thus, the objective of this last step of the proposed methodology is divided in two differentiated classifications:

1. Classification between phytoplankton specimens and the rest of the candidates. It aims to reduce the amount of false positives that the previous step let through. It is done by creating two classes, one contains all the phytoplankton specimens and the other class the rest of specimens. This class groups zooplankton and garbage among other things.
2. Identification of target phytoplankton species. It intends on separating the different species of phytoplankton with special focus on those that are a risk to public health. It uses the phytoplankton specimens from the previous step. It should be noted that, while this classification should be the continuation of the previous one, it was chosen to test it with a dataset made out of the true phytoplankton classification, that is, without noise. This was done to evaluate the performance of the system without interference from any potential noise that could come from the previous classification.

Both classification steps are approached using a common classification methodology. This



(a) Specimen

(b) Binary mask

Figure 5.1: Specimen from the dataset and its corresponding mask

methodology consists on, first, the representation of the phytoplankton specimen candidates using a set of features and the classification of these candidates in the defined feature space. The following sections are used to describe this classification methodology. In section 5.2 the dataset of specimens from the previous chapter is described. Section 5.3 is devoted to the used feature representation, while section 5.4 explains the different classifiers used and 5.5 focuses on the metrics used. The chapter is then split into the two classifications, phytoplankton detection 5.6 and species classification 5.7. Both sections contains subsections detailing the experimental setting, results, discussion an conclusions of each one.

5.2 Dataset

The dataset used in this step is composed of the candidate detections that resulted from the previous chapter. This includes a bounding box that encloses each candidate and the image mask used to produce it, an example of such pair can be seen in figure 5.1. It should be pointed out no discards where made, that is, all the false positives from the system are kept to continue with the trend of real world usage as well as to develop a beginning to end system.

The only notable change from the previous set of images to this is the fact that to enforce a policy similar to the one used by biologists the detections whose bounding box touched an edge, that is, was incomplete, were left out. An example of these types of candidates that were left out can be seen in figure 5.2. Currently there is no fixed policy for discarding incomplete organisms in the manual analyses, but it is up to the criterion of the expert. This varying criterion worsens repeatability of the analysis and inter-expert consensus. By enforcing a clear policy for these discards the system output will always be consistent.

Removing the incomplete specimens from the dataset should also ease the job of classifiers, adding incomplete organisms might result in worsened performance due to the impor-

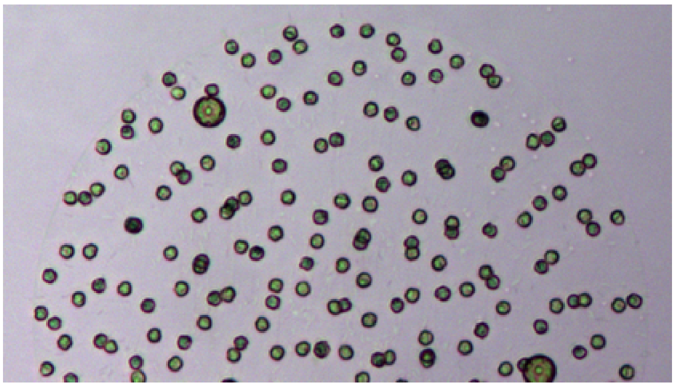


Figure 5.2: Type of discarded specimen due to its incompleteness

tant missing information.

The training and test split is kept the same as in the dataset in the previous chapter. This means that all the bounding boxes detected on the training set images will be used to train, all the ones detected on the test set images will be used for the testing.

The ground truth varies for each of the classifications that will be done in this step, as each bounding box receives a different tag depending on the intention of the classification. This means that, while the data stays the same, the tag of a specimen will be phytoplankton or non-phytoplankton for the first classification while for the second one it will be its concrete species like *Woronichinia naegeliana* or *Dinobryon Sp.* It should be noted these classifications have been agreed upon with an expert biologist, as the taxonomic classification of these specimens requires a lot of training and expertise.

The images, in this step are not homogeneous like in the previous one. Every specimen has a different size, which means that every image has a different resolution and aspect ratio matching the detection it encloses. Each specimen is accompanied by a binary mask that allows the separation of the specimen itself from the background.

In the figure 5.3 several examples of specimens with their corresponding masks can be seen. Each specimen has its mask on its right so they can be easily spotted. The organisms are separated according to the classifications that will be done in this step. The outer-most class, that referring to a specimen being phytoplankton or not, corresponds to the first classification which aims to clear the noise from the detections from the previous step. The second classification corresponds to the inner-most types shown in the figure, that is, the species. It should be noted that the specimens included in each group, specially those of concrete species were chosen to represent most of the morphological differences between specimens belonging to the same species. It is also important to mention that in the dataset no specimens of *Microcystis flos-aquae* were present despite the fact that they are common cyanobacteria in the Galician lakes. All the images of the specimens, except for those on the *M. flos-aquae*, are

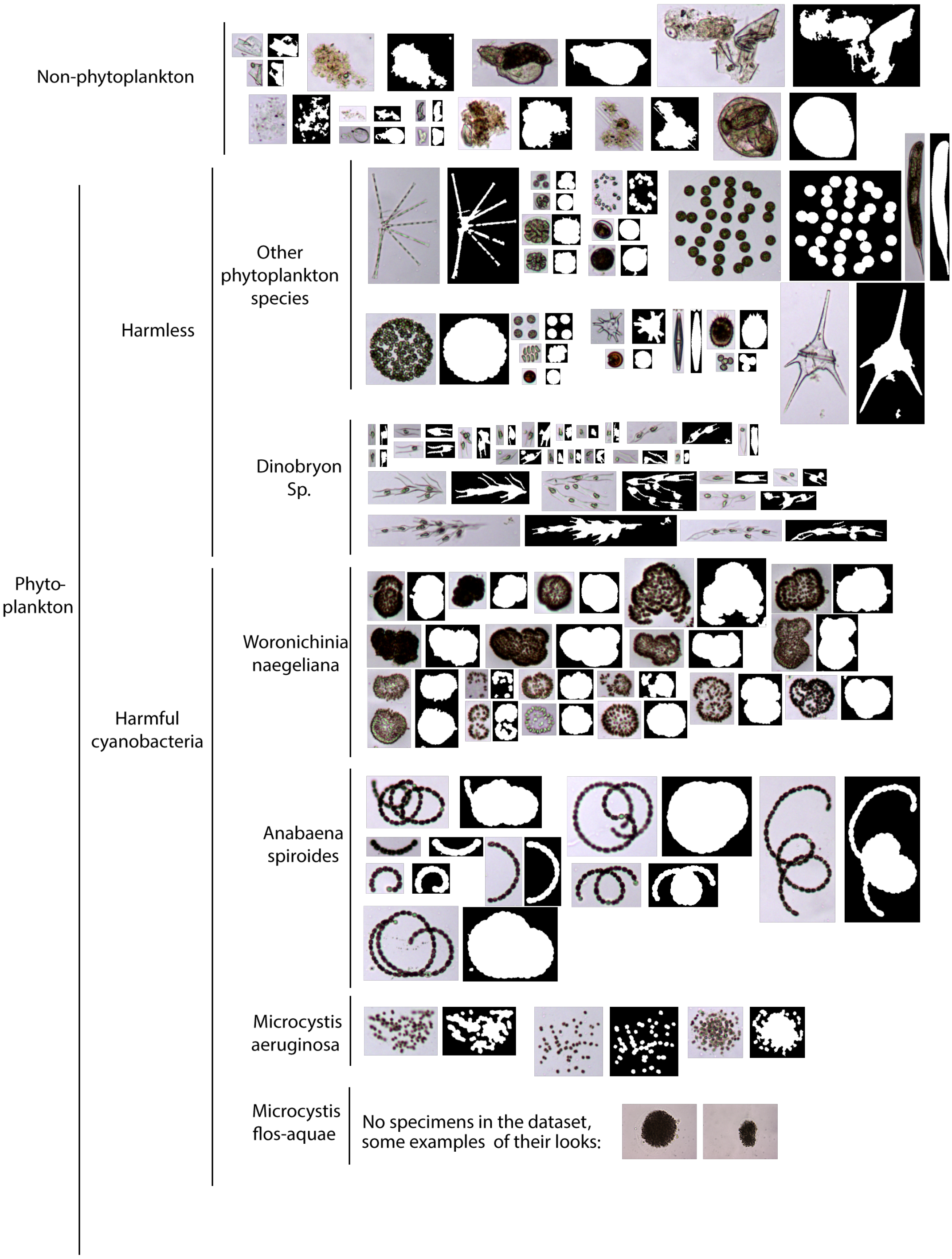


Figure 5.3: Example of all the specimen types that make up the dataset

to scale in relation to one another. This means that their relative size is the same as it is in the dataset. The differences in size can be specially notable in *Dinobryon Sp.* or in *Anabaena spiroides*. Overall, the dataset holds images of very varied sizes from the smallest with a size of 11×13 pixels to the biggest, measuring 314×699 pixels. The masks that accompany each specimen have the same size as the specimen.

When creating the tags for the second step of the classification a problem emerged. Originally, the plan was to classify the specimens in the dangerous species and a rest or harmless group. That, with the types of dangerous cyanobacteria present in this lake, would mean a total of five groups. Out of the total four species of cyanobacteria, only one was present in a number sufficient for a classifier to learn, *Woronichinia naegeliana*. The other three *Anabaena spiroides*, *Microcystis aeruginosa* and *Microcystis flos-aquae* were too scarce in the dataset for a classifier to properly work.

After this setback a new set of relevant classifications was devised to test the system in different aspects and, at the same time, fulfil the original intention of doing a potability analysis. The new set of classifications are:

- ***W. naegeliana* vs the rest**: The most abundant of the dangerous cyanobacteria and the only one numerous enough to be classified alone. It is, therefore, the ideal case to classify in single species versus the rest fashion.
- ***Dinobryon Sp.* vs the rest**: The most common phytoplankton organism in the dataset, it is not harmful as it does not produce toxins. It has several distinctive features but it also has a very high variation between specimens that makes it a challenging problem to classify. It is an ideal case to benchmark the system capabilities.
- **Harmful vs harmless**: The scarcity of some dangerous species does not allow to classify them alone, however, grouping them all up does so. This is, perhaps, the most important classification, in terms of potability testing. It allows to count the dangerous organisms to produce a metric in the style of cyanobacteria per millilitre required to know if the water is safe. It is similar to the original idea of classifying between dangerous species and the rest but it adds complexity as the system must generalise the features that distinguish the cyanobacteria as whole from the harmless phytoplankton.

5.3 Feature extraction

The features used in this work to represent each of the images in the dataset are divided into three main groups: texture, colour and morphological. Both the texture and colour features are based on the Bag of Visual Words model applied to different base spaces. A space of Gabor filters is defined to describe texture while the RGB space is used to describe colour. The morphological features, instead, are based on the shape and size of each specimen.

5.3.1 Bag of Visual Words model

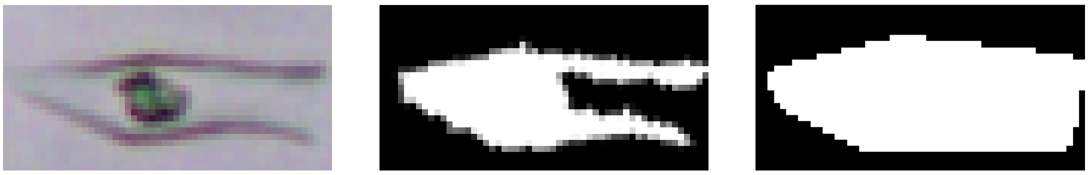
The Bag of Words model [69] was born in the information retrieval and natural language processing scopes, a more in-depth explanation can be found in appendix A.8. Since then been successfully adapted into the Bag of Visual Words (BoVW) used in computer vision. This model uses a set of visual words which is a visual feature or patter that may be present or not in the images to analyse. A typical procedure to compute the dictionary of visual words consists on gathering descriptors for several feature points in the training images. These points are visually located at invariant keypoints in the image (obtained trough SIFT [70] or SURF [71] algorithms) although dense sampling can also be used. Once the local shape information of the image at each of these points is described using a feature vector, all the point samples are clustered (using, e.g. k-means) to identify a set of visual words (one for each cluster centre). The Bag of Visual Words descriptors of an image or region, consist on the histogram along these visual words. The representation represents the visual words statistics of the region, which effectively accounts for the balance between the different local shapes.

In this work the algorithm used to obtain the feature space of the visual words is not any of the described above but Gabor filter banks. By applying a set of filters to the image a vector of texture features is computed fore each pixel. This space is clustered to obtain a set of visual words for which histograms are computed. This poses an implementation of the same concept using a densely sampled feature space using Gabor filter banks.

The process followed by the BoVW can be seen in B.5 and is summarised as follows:

1. **Clusterise the information:** All the feature vectors for the images which in this case encode local shape through a Gabor filter bank, are clusterised using K-means. This allows a reduction o the information as it condenses similar features into visual words. The number of centres (k) is a free parameter, which best value can be found through testing. This step creates a dictionary model, containing the cluster centres (visual words) that will be used further in this process.
2. **Create a histogram:** The dictionary of visual words is then used to create a histogram. Each visual word represents a bin in the histograms, thus the histogram accounts for the number of times that the corresponding pattern appears in the represented image region. Once the histograms are calculated they are normalised to obtain a probability. A version without normalisation will be tested as well. These histogram vectors are finally used as descriptors for the image regions.

These two steps represents the training and simulation part of the bag of words model. The training is performed using the information of the training set of images. The simulation, instead, can be performed with both the training and test images to compute the training and test features vectors respectively. These feature vectors are the ones that will be tested with different classifiers to evaluate their performance. It should be noted that, to compute the



(a) Original specimen

(b) Original mask

(c) Convex hull mask

Figure 5.4: Demonstration of the mask filling algorithm

colour features, the same process is followed changing the usage of the Gabor filter banks to the RGB information, while the rest remains the same.

The process of computing the feature space for this model, be it using Gabor filter banks or RGB information, is subject to two different implementations: masked and unmasked. This relates to the previously described specimens that each included their mask. The unmasked version computes the feature space over all the candidate images, that is, the bounding box. In this case the image includes background data. It is hypothesised that such data may skew the data and worsen the features but it could also include environmental and external features to the specimens that might help in its classification. The masked version takes only the pixels that are inside the mask for each specimen, that way the amount of background in the computations is minimised. Both methods will be tested in the project. The only change made to the masks has been to fill the holes that some of these mask have, with the intention of providing information that might have been, otherwise, missing. This would allow to capture information about the contents of some transparent or sparse specimens like in the case of *Dinobryon Sp.* which masks tend to miss its capsule as can be observed in the figure 5.3 due to its transparency. This process was done using a convex hull that encloses the white blob of the mask, effectively filling holes that it may have inside as well as attaching any part that may not be connected to any other. In the example provided in figure 5.4 a *Dinobryon Sp.* whose original mask would miss part of its capsule. Using the convex hull the the new mask is able to fill the missing spots in the previous version and therefore should be able to contribute its full information to the algorithm.

The free parameter left in this algorithm, k, is the amount of words that is calculated. In the testing step the methodology will be evaluated with a total of eight amounts of words: 100, 50, 20, 10, 8, 5, 3, 2.

5.3.2 Gabor filter banks

Gabor filters perform local Fourier analysis and are essentially the product of a Gaussian kernel and cosine and sine functions. Each of these filters is only sensitive to a given frequency and orientation of the signal which, basically means that the filter checks whether there is any specific frequency in a specific direction, in each location of the signal. They are widely used and regarded as useful tools to extract texture features .

The Gabor filters are composed of parallel excitatory and inhibitory stripe zones, this can be appreciated in the figure 5.5 as black and white stripes. The shape of these stripes is controlled by a parameter called ellipticity. It controls the aspect ratio of the strips, changing it from straight to elliptical.

Gabor filters are applied over the image using a 2D convolution. These filters accept several parameters that tune the response they get out of the image. The parameters are:

- Orientation: the Gabor filter can be rotated with angle θ, this changes the region of the image that the filter is sensitive to. Usually in the Gabor filter banks several versions of the same filter are tested with different angles of rotation with the intention of checking different directions. To obtain responses in all directions at least 4 different orientation are needed like 0°, 45°, 90° and 135° but more granularity and detail can be extracted with more orientations. This parameter will be tested with 4 and 8 orientations in the filter banks for this project.
- Bandwidth: the smaller the bandwidth is the bigger the width of the of the Gaussian envelope used in the Gabor filter gets. This effectively controls the amount of pixels that the functions analyses as it increases the amount of stripes that the filter creates. Four different bandwidths will be tested.
- Frequency: this parameter controls the spatial frequency of the harmonic function, this modifies the wavelength and thus the amount of pixels that each stripe of the filter analyses. Twelve unique frequencies corresponding to semi-octaves, will be tested.

The effects of this parameters can be seen the example in figure 5.5 as they change the output image. Trough these parameters a series of filters can be tuned in order to obtain the most desirable output for the set of images to analyse. Several configurations will be tested along this work in conjunction with the bag of words. This three parameters already net a total of 96 unique combinations of filter banks that need to be tested in conjunction with the different settings for the BoVW

$$g(x,y;\lambda,\theta,\phi,\gamma) = \exp\left(-\frac{x'^2+\gamma^2 y'^2}{2\sigma^2}\right)\cos\left(2\pi\frac{x'}{\lambda}+\phi\right), \tag{5.1}$$

Where λ is the frequency, θ is the angle, ϕ is the phase and γ controls the ellipticity. The x' and y' coordinates are rotated space according to:

$$\begin{aligned} x' &= x\cos\theta + y\sin\theta, \\ y' &= -x\sin\theta + y\cos\theta. \end{aligned} \tag{5.2}$$

Where θ denotes orientation

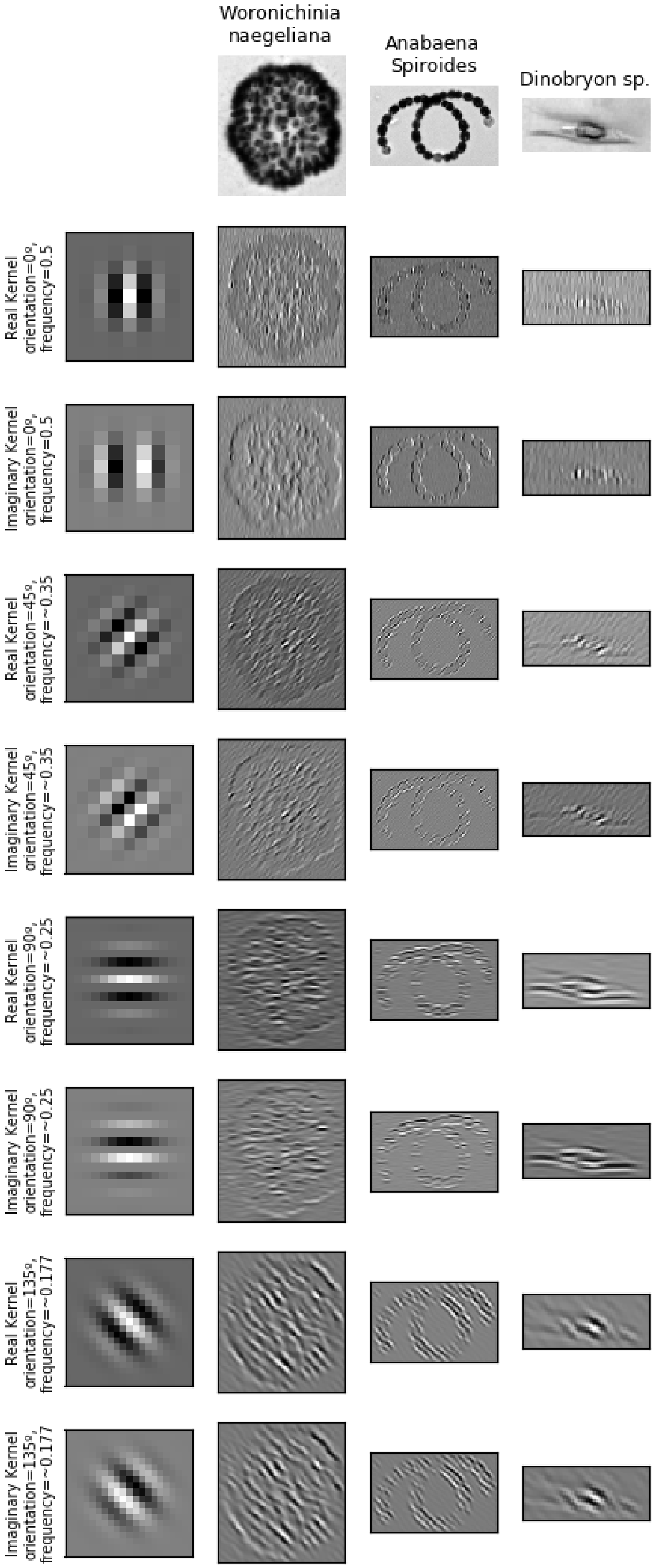

Woronichinia naegeliana
Anabaena Spiroides
Dinobryon sp.
Real Kernel orientation=0º, frequency=0.5
Imaginary Kernel orientation=0º, frequency=0.5
Real Kernel orientation=45º, frequency=~0.35
Imaginary Kernel orientation=45º, frequency=~0.35
Real Kernel orientation=90º, frequency=~0.25
Imaginary Kernel orientation=90º, frequency=~0.25
Real Kernel orientation=135º, frequency=~0.177
Imaginary Kernel orientation=135º, frequency=~0.177

Figure 5.5: Example of Gabor filter bank responses

5.3.3 Colour information

The colour information is computed the same way as the texture data, through the bag of visual words model, but instead of using Gabor filters as the feature space it uses the RGB data. The process followed by the algorithm is exactly the same, clustering the data obtained from each of the separate RGB channels to form visual words. Again the parameter k is free and will be found through testing and experimentation. The histogram for each of the words is calculated like previously as well as normalised or not depending on the test. The only difference is that two different versions will be tested, one of them varies slightly the process as it tries to join both texture and colour in the same feature set. The two different methods are:

- **As another bag of words instance:** The BoVW works exactly the same as it normally does, just over the RGB feature space. In the end the feature vector obtained will be appended to the obtained from the texture bag of words. This method computes the texture and colour features separately.
- **In the same bag of words as Gabor:** This version fuses the texture and colour information before the k-means clustering of the bag of words. This is done since both colour and texture are complimentary data and k-means will be able to cluster the visual words normally.

  Comparison of both will be done in the results section

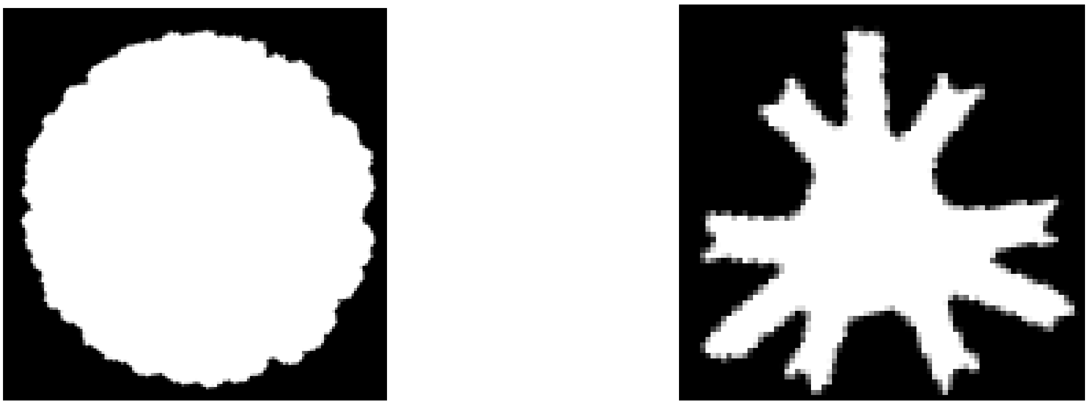
It should also be noted that colour information will be tested, like Gabor filters, with unmasked and masked channels. The main difference is that the masked approach only takes into account the pixels that are inside the mask discarding the rest while unmasked uses all the image. The masks are exactly the same as those described previously, in section 5.3.1.

Like in the previous explanation of the BoVW algorithm, with colour information the same 8 quantities of words, 100, 50, 20, 10, 8, 5, 3 and 2, will be tested to find out which one is the best performing.

5.3.4 Morphological information

Morphological information can also be an interesting feature to distinguish phytoplankton. It is however, difficult to obtain a reliable shape or size descriptor since phytoplankton is defined for its plasticity and variability. Two different descriptors are, then, devised in order to test their performance in conjunction with the rest of the features previously mentioned.

- **Amount of pixels in the mask:** The amount of pixels in the mask is simply counted and used as feature. It acts as a descriptor of the size of the particular specimen as the masks are proportional one to another. This feature could prove useful as some species have very distinguishable sizes, like *Dinobryon Sp.* that tend to be small or *Volvox Aureus*



(a) Rounded specimen

(b) Irregular specimen

Figure 5.6: Comparison of masks for a regular and a irregular specimen

that are often very big.

- **Proportion of pixels in the mask:** Similar to the previous one, the number of pixels in the mask is counted but it is divided by the total amount of pixels present in the mask image. This results in a proportion of specimen size versus minimum enclosing rectangle size. This serves as a shape descriptor in the sense that a round, symmetrical specimen will have more proportion of white pixels than an irregular specimen that required the bounding box to be enlarged due to its size. A visual example of this is provided in figure 5.6 that compares a regular specimen 5.6a to an irregular one 5.6b.

5.3.5 Summary of the tested features

Overall, three groups of features will be tested: texture, colour and morphology. The texture and colour features are based on the application of a Bag of Visual Words with Gabor filter banks and RGB channels respectively. The morphology features intend to be a descriptor of shape and size. They use the number of pixels in the mask or the proportion of pixels in the mask when compared to the whole image to obtain data about the overall shape and size of the specimen.

The Bag of Visual Words creates a dictionary from all the visual words present in the training set by clustering them. These features come from the selected feature spaces, in the case of the texture Gabor filter banks and in the case of colour RGB channels. Both of these BoVW will be tested with 8 different amount of clusters. It should also be noted that, each one of those, will be tested with masked an unmasked images. The unmasked images provide information of the background while the masked ones do not. It has been hypothesized previously on favour of using each one, so testing both is necessary. This effectively duplicates the amount of testing needed.

Gabor filter banks are used to create the texture feature space used later in the BoVW. Each of these filters is sensitive to a special frequency and orientation of the signal present on the image, this means that these filters can be highly tuned to fulfil the need of any work. A total of 96 unique filter banks are tested together with the 8 previous amounts of clusters. This makes 768 possible combinations, but the usage of masked and unmasked images duplicates the amount, up to a total of 1536 tests needed just using texture information.

The RGB channels are used to extract the colour features through the BoVW, similarly to the texture features. To combine these features with the texture ones two different methods are devised: using two separate Bag of Words and mixing both features in the same Bag of Words. Both of these approaches will be tested to see which one performs better. It should also be noted that, again, 8 different quantities of clusters will be tested to find out which is the one that shows better performance.

The morphology descriptors try to measure and the shape and size of the phytoplankton individuals. Two different descriptors were created. The first one just counts the white pixels in the mask, effectively quantifying the size of the specimen. The second descriptor also counts the white pixels in the mask, but it divides their quantity by the total amount of pixels in the mask image. That way it gets a proportion of the pixels in the mask, which act as as a shape descriptor since more irregular specimens will have a lower proportion and more regular, rounder ones will have a higher proportion of white pixels.

5.4 Classifiers

Several classifiers are tested using the different features explained above. The aim of this step is to find the classifier with the best performance for the images at hand and the chosen features. The five selected set of classifiers is detailed below, more detailed information can be found in section A.8 of the first appendix.

- **Support Vector Machines:** Separate the data arranged in an n-dimensional space using hyperplanes that maximise the classification margin. The kernel variations, furthermore, can result in complex classification boundaries. The kernel variants can, however, suffer from the curse of dimensionality when using large feature spaces.
- **k-Nearest Neighbours:** Groups the data basing it on a point's similarity to its neighbours and a similarity function that can be adapted to the target problem. Several similarity metric will be tested to find out the more suitable one. It is included in here as it can asses if there are clusters in the feature space with samples of the same class.
- **Boosting Trees:** Groups decision trees using a boosting algorithm to create a strong committee classifier. Apart from the cascade classification of the boosting algorithm, the trees are able to develop complex classification boundaries on large feature spaces.

- **Random Forest:** Uses bagging of decision trees to create a strong classifier in which each tree focuses a subset of features. Random Forests are able to develop complex classification boundaries.
- **Gaussian Mixture Models:** Models the distribution of the feature space for each class following a mixture of Gaussian distributions of unknown parameters. These distributions can be then used to classify samples using the Bayes Rule, which can develop complex and smooth classification boundaries.

Each one of these classifiers has several parameters that are tuned to find the best performance over the training set. This is done through grid search, which allows to test the selected set of parameters. The amount of configurations to test increases by orders of magnitude as some of this classifiers have many parameters that must be tuned.

5.5 Evaluation metrics

Both, phytoplankton detection and species classification, are optimised toward the same metric, maximise the precision at a given minimum recall.

To correctly define the metrics, it should be first established what several terms mean for each one of the two classifications:

- **True positive:** In the first classification it is a phytoplankton specimen correctly labelled as such while in the second it is a specimen belonging to a harmful species correctly classified in its species.
- **False positive:** It could be expressed as a false alarm. It means that a non-phytoplankton specimen is included in the phytoplankton group in the first classification. In the second one it would mean that a harmless individual is included in the harmful species.
- **True negative:** A correctly discarded specimen, corresponds to a non-phytoplankton correctly labelled as so in the first step and a harmless specimen correctly identified.
- **False negative:** A member of the positive class is incorrectly classified. This means that a phytoplankton is labelled as non-phytoplankton or that a harmful specimen is classified as harmless. In this work, this is the more dangerous error to make as it means that interesting or dangerous (respectively for each classification) specimens are misclassified.

Precision or positive predictive value is the metric that describes the set of samples correctly labelled as positive divided by all the positively labelled samples. The equation 5.3 describes this metric.

$$recall = \frac{True\ Positives}{True\ Positives + False\ Positives} \tag{5.3}$$

Recall, also commonly called sensitivity or true positive rate is a metric that is calculated dividing the amount of elements positively identified between all the positive samples. That is, it intends on evaluating how many of the relevant items are correctly labelled. It can be described with the equation 5.4.

$$recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \tag{5.4}$$

Both of this metrics are combined in the one that is used to evaluate the systems. Precision at a high level of recall intends to evaluate the quality of the selected specimens when most of them are selected. The recall levels used are 90%, 95% or even 99%. This metric is selected as the focus of the work is ensure that the bare minimum of the interesting specimens are misclassified. In both classification this means that the positive class, be it phytoplankton specimens or dangerous specimens, needs to contain nearly all of the specimens truly in that class in order to avoid an skewed analysis. The most important misclassification, as it can be derived from the metrics used, is the one that puts a positive class sample into the negative one. For each classification it means assigning a phytoplankton specimen to the non-phytoplankton group or a dangerous specimen to the harmless class, respectively. The usage of precision at a high level of recall is, then, orientated towards real world usefulness.

This metric can, also, be plotted in a precision-recall curve. This type of curves plot both metrics, one in each axis, which allows to see how the precision fares at different levels of recall. Due how they are drawn, the ideal classifier would reach the point (1,1) meaning that it obtains 100% precision at 100% of the recall, the perfect classification. The closer the curve gets to that point the better the classifier is.

5.6 Phytoplankton detection

This section tackles the classification of the candidate detections obtained with the methods described in chapter 4, into two different classes: phytoplankton and non phytoplankton. This aims at filtering all the detections that are not phytoplankton like garbage or zooplankton.

First, the experimental setting is described in 5.6.1 section. The detailed results are provided in section 5.6.2. Section 5.6.3 is dedicated to the discussion of the results previously mentioned. Finally in the section 5.6.4 a brief summary of the previous points and their main conclusions are given.

5.6.1 Experimental setting

The experiments will be done over the previously described dataset focusing on maximising the precision at a high level of recall.

To stabilise the system and get more consistent results it was decided that, during the

Number of bins/clusters	100,50,20,10,8,5,3,2
Bandwidth	2,1.5,1,.05
Frequency	0.5, 0.3535, 0.25, 0.177, 0.125, 0.088, 0.0625, 0.0442, 0.03125, 0.0221, 0.0156, 0.01105
Orientations	4, 8
Masked	unmasked or masked
Colour features number of bins/clusters	None or 100,50,20,10,8,5,3,2
Shape features	None, number of pixels in the mask or proportion of pixels in the mask

Table 5.1: Different Gabor and bag of visual words settings used in the tests

training of the system, k-means in the Bag of Visual Words would be run 10 times. After that the best one according the chosen metric over the training would be chosen to continue further. This allowed to obtain more consistent results. After all of that, some randomness is still present on the system. That means that if all the tests were ran again the results will likely not be the same although they would be quite similar and in a acceptable range.

As already mentioned previously the amount of unique configurations needed to test is really high. It was decided to minimise the combinations by adding the colour features to the top three best performing texture features. The colour features are also tested separately and the three best results are gathered to be fused with the previously mentioned top three of the texture features. Then the shape features will be tested with the best combinations of those two. This dramatically reduces the amount of tests.

Both Bag of Words and Gabor filters have several parameters that need to be tweaked to improve the results. Several parameters of the Gabor filters can be altered to produce different outputs, as it was mentioned previously. This settings together with the ones from the Bag of Visual Words are detailed in the table 5.1. The combination of all the texture features results in 1536 unique tests. With the top three results the colour features will be tested. Separately, the colour test amount to a total of 32 test. After selection the top three of each they are combined in all possible ways which means 9 further tests. Finally, the two morphological descriptors are added to the top three of the previous combination, separately. That means 6 extra tests, to finalise with a total of 1583 tests per classifier, without taking into account the optimisation of the classifiers' parameters.

Each classifier has different number of parameters that need to be adjusted. Using grid search said parameters are adjusted, finding the ones that offer the best performance over the training set. The amount of tests needed per classifier given their amount of tests is explained in the following list

- SVM: Due to its low amount of parameters, SVM only 52 unique tests are needed. With the previous calculation of the feature combination this amounts to 82.316 tests.

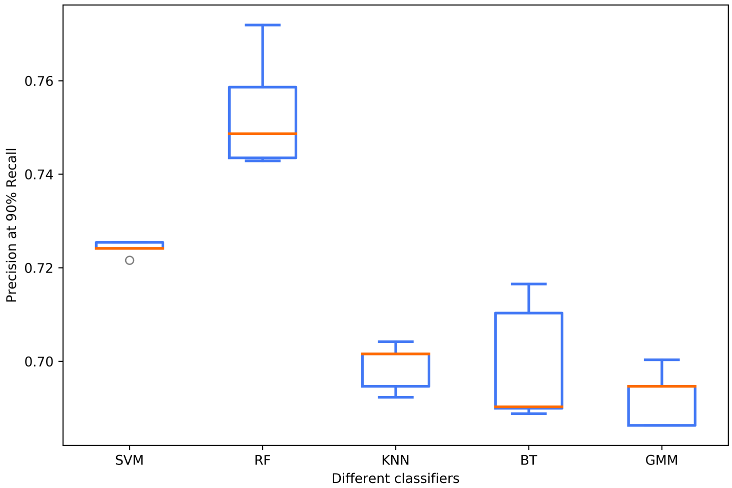


Figure 5.7: Box plot of the 5 best performing classifiers for each kind

- k-NN: Also required low amount of tests, a total of 48. This results in a total of 75.984 tests to be done.
- RF: One the most test-heavy classifiers due to it being highly configurable. 120 tests were done for this classifier which amounts to a 189.960 unique combinations tested.
- BT: Like RF the amount of parameters this algorithm can adjust is very big so it required the most test, 162. The total amount of configurations for this classifier, with the data from the the features, is then 256.446.
- GMM: It requited the lowest amount of tests, a total of just 8. The total amount ascends to 12.664 unique combinations of parameters and features.

The total amount of tests done in this classification is 617.370. This prompted the necessity of using the servers detailed in the section 3 that allowed for a fast execution of such a huge amount of tests.

5.6.2 Results

Due to the high amount of tests made for each classifier, as stated in the previous section 5.6.1, it was chosen to present the results of the best performing for each one only. This way the clutter derived from thousands of results, some of them irrelevant, is avoided.

The results for the classification of the detections into phytoplankton and non-phytoplankton measured by their precision at 90% of the recall are presented in the figure 5.7. The top result for the whole set of experiments is obtained by a Random Forest classifier showing a 77% of precision at 90% of the recall. RF is overall the best performing classifier with the rest of the classifiers lagging behind in terms of performance. Only SVM manages to show comparable performance to random forest.

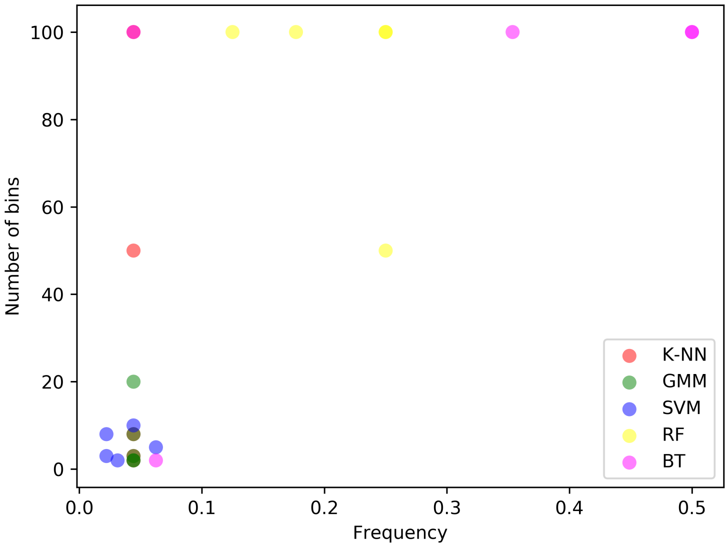


Figure 5.8: Scatter plot of the number of visual words and Gabor filter frequencies for the best 5 performing classifiers of each type

In figure 5.8, a scatter plot showing the number of visual words and the Gabor frequencies for the best performing classifiers of each type. It shows which parameter each classifier benefits from. Two main groups emerge, on one side, BT and RF favour large amount of bins not caring about the frequency while the rest, k-NN, SVM and GMM favour smaller frequencies as well as a low number of bins.

In the figure 5.9 the Precision-Recall curve for the best performing classifier can be seen. It is a Random Forest classifier that used 100 visual words of Gabor filters, that is, just texture information. The frequency for the Gabor banks used is 0.25, with 8 orientations and 1.5 as the bandwidth.

Figures 5.10, 5.11 and 5.12 show scatter plots of the RF classifier performance (in terms of precision at 90% of recall) with varying Gabor filter and BoVW parameters. Figure 5.10 shows the performance obtained with varying number of visual words for the best performing configuration. Figures 5.11 and 5.12 show similar graphs with varying values of bandwidth and frequency, respectively, for the Gabor filter banks.

Tables 5.2 and 5.3 show different runs of the top classifier for each type, respectively. It can be seen that, the standard deviation, for precision at any point of the recall is really low.

The colour features by themselves managed to attain several high precision results but never were able to surpass the texture only classifiers. The best results for colour only remained at 71% while the best for texture only was of 77%.

To compare methods of joining the colour features to the texture ones is a little bit different as the advantage that using different bag of words has is that it can feature two different amount of visual words for colour and texture whereas using a single bag implies that the

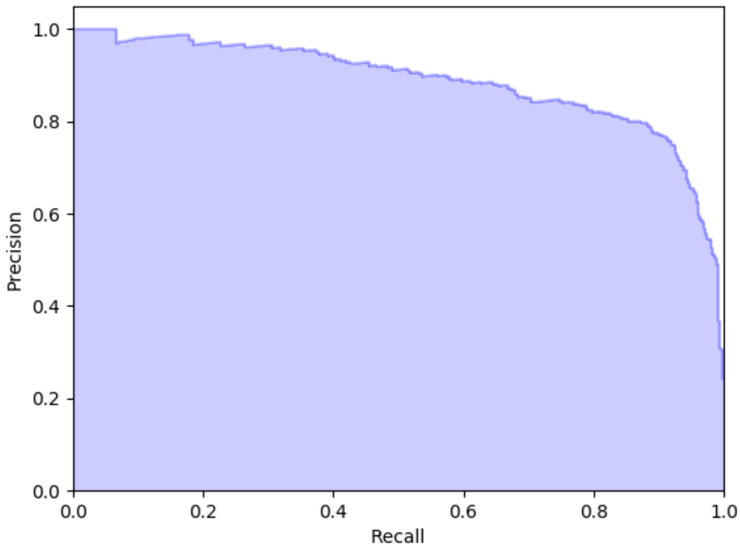


Figure 5.9: Precision-Recall curve for the top classifier, a RF with 100 visual words and Gabor filter banks with 8 orientations, 0.25 as frequency and 1.5 as bandwidth

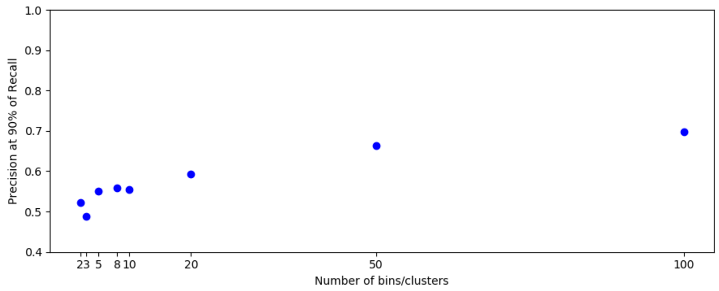


Figure 5.10: Results for the best performing classifier (RF) varying the number of visual words

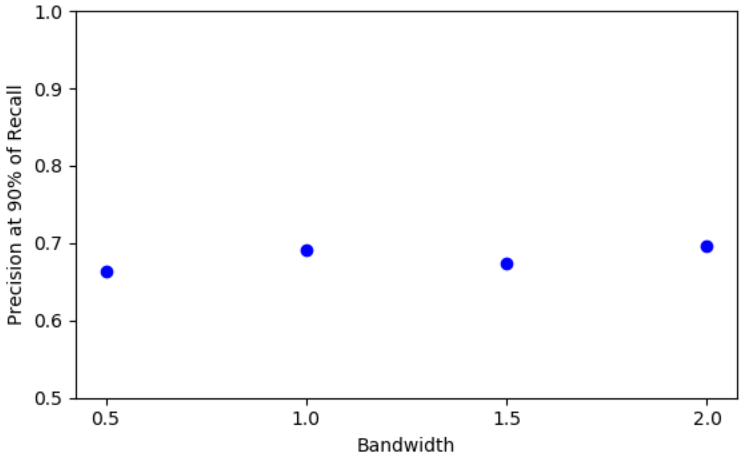


Figure 5.11: Results for the best performing classifier (RF) varying the bandwidth of the Gabor filter banks

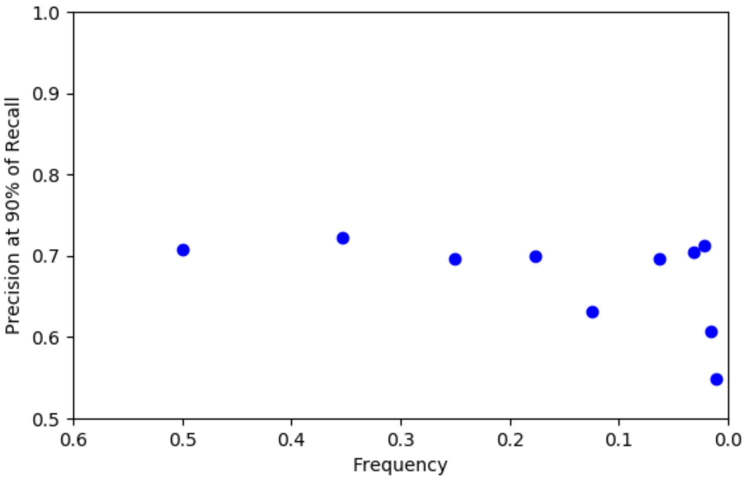


Figure 5.12: Results for the best performing classifier (RF) varying the frequency of the Gabor filter banks

	Precision at 90% of Recall	Precision at 95% of Recall	Precision at 99% of Recall
Run 1	0.722	0.611	0.376
Run 2	0.72	0.597	0.419
Run 3	0.737	0.631	0.341
Run 4	0.733	0.652	0.449
Run 5	0.748	0.603	0.442
σ	0.01	0.02	0.041

Table 5.2: Five different runs of the best Random Forest classifier

amount is shared. Comparing the best performer of each group yields a significant different between them, with the two separate bag method pulling ahead in with 73% of precision at 90% of the recall while the same bag method shows a top performance of 68% of precision at the same level of recall.

Another tested metric was the normalisation or not normalisation of the histograms. Again, using the top performers for each category, the difference is substantial. With normalisation the top result is the already mentioned 77%, without normalisation the best result is 66.5% of precision at 90% of the recall.

The unmasked feature showed better performance with the top result of 77% while the highest precision at 90% recall of the masked ones were 73.8%. The top five results for each can be visualized in figure B.1 located in the second appendix.

5.6.3 Discussion

Many different settings and parameters were tested for each one of the five classifiers. Early on in the experiments, the testing of normalisation versus non normalisation regarding the

	Precision at 90% of Recall	Precision at 95% of Recall	Precision at 99% of Recall
Run 1	0.735	0.622	0.319
Run 2	0.719	0.685	0.37
Run 3	0.731	0.618	0.318
Run 4	0.728	0.697	0.334
Run 5	0.725	0.685	0.34
σ	0.0054	0.0341	0.0189

Table 5.3: Five different runs of the best support vector machine classifier

histograms in the Bag of Visual Words revealed that normalising the histograms showed a significant bump in performance. This has already been mentioned in the previous section and can be seen in the example provided by above. From there on, the tests were made solely with normalised histograms which effectively halved the number of tests greatly reducing work. It should be noted that, on the count made previously this was already not taken into account. Not having done this reduction of tests, the total amount for this step would have surpassed the million.

When combining colour features with texture, the approximation with two different bags proved to perform better than the single bag approach. This was exemplified previously, on the previous section 5.6.2. The difference in performance promoted that, for next step this tests would not be carried out, only the two-bag approach. It is done to reduce the big amount of tests as the ones in this classification took a long time to execute, even with full access to two servers.

Regarding the tests done with masked and unmasked settings, the unmasked one showed consistently superior performance as can be seen in figure B.1. It is hypothesised that this difference is likely due to the fact that not masking the image may lead to some shape or size extraction within the texture features that could prove beneficial for the classifiers.

In terms of features, adding colour or shape features to the texture ones proved to be detrimental to the performance of the system, as it was said in 5.6.2. This is likely due to the fact that both classes are completely heterogeneous, specially the non-phytoplankton one. Many garbage and zooplankton specimens have similar colours to the real phytoplankton specimens which may cause the reduction in performance. The morphological descriptors are affected by this as well since many phytoplankton organisms do not have firm, characteristic shape or size and neither do the non-phytoplankton elements. All in all, the intraclass differences and interclass similarities seem to be more exaggerated with shape and colour features which may cause them to worsen the performance of the texture-only classifiers.

All in all, it is concluded that both the Gabor filter parameters and the number of visual words does prove to be of major importance in the classification. This can be appreciated in the results shown in the figure 5.10, for the particular case of the best performing Random

Forest classifier. In the case of SVM or any other classifier, while the pattern may be different, the take away is the same. Results are heavily influenced by the settings chosen and the ones that benefit the classifier the most. The influence of the various settings can also be viewed in the figures 5.10, 5.11 and 5.12. The differences between the results may not seem that big, but it must be taken into account that in each one of them only only a particular setting is being modified thus each of the parameters matters.

The randomness present on the system while it is something to take into account it is not too big or unaccounted for. This can be seen in tables 5.2 and 5.3. Both methods show more or less the same deviation of about 1-2% between those five results. This bigger discrepancy between results could be attributed to a bigger difference between runs or simply a outlier in the performance of this method. The bigger discrepancy could be traced back to the fact that RF has many more parameters that grid search can specifically tune. This higher number of hyperparameters may cause the algorithm to focus on noise or overfit, which coupled with the small changes that k-means introduces in the bag of words could explain the variation in the results.

With all the results laid out, it is important to note the high difficulty of this particular step. Both classes have big amounts of variability in them since they are grouping a very broad spectrum of detections. The positive class, phytoplankton, groups very different organisms with different colours, textures and shapes. The variation in the non-phytoplankton class is even bigger as it groups alive organisms like zooplankton and other detections like inorganic garbage. This adds a great amount of difficulty to this step but it is overcame with good results by the classifiers, specially RF and SVM. To further reduce the amount of tests needed, in the next classification step, the one that intends to separate dangerous species from the harmless ones, only SVM and RF classifiers will be tested as they proved to have the higher performance.

5.6.4 Summary and conclusions

This section the classification of candidates into phytoplankton and non-phytoplankton candidates (including zooplankton and garbage) was approached using several alternative classifiers and feature sets. The chosen features were texture, using BoVW over Gabor filter banks, colour using BoVW over RGB and, finally, morphology through counting and proportion of the pixels in the mask.

Five different classifiers were tested using these features varying the parameters of said features and classifiers. The chosen metric was the precision at 90% recall to minimise the misclassified phytoplankton organisms.

The top result yielded a 77% of precision at 90% of the recall using a Random Forest classifier that used only the texture features. In this case, combining these features with colour resulted in no benefit. Using only colour produced results nearing that of the texture features but it was never able to surpass it. Shape features turned out to be useless in this step most likely due to high variation in the classes.

Overall the main conclusion drawn, is that, despite the high difficulty of this step most of the classifiers and feature sets are suitable for reaching an acceptable precision at 90% recall for the application.

5.7 Species classification

This section tackles the classification of phytoplankton specimens into different species. The aim is to differentiate the dangerous species that produce toxins from the harmless ones in order to use those results for a potability test of the water.

First, the experimental setting is described in 5.7.1, by detailing how the tests where prepared and the details of them. Results are presented in 5.7.2. This lays out the data of corresponding to those experiments. Later, in the section 5.7.3 those results and their meaning are be discussed. Finally, in section 5.7.4 a brief summary of this experimental process and its main conclusions are given.

5.7.1 Experimental setting

This step follows the same procedures as the previous one as well as the metrics that the system was optimised for. As it is mentioned on the previous sections, the amount of tests is reduced discarding some features that proved to perform poorer than others as well as some classifiers, for the same reason. This means dropping the same bag colour and texture features, the non-normalisation of the histograms as well as three classifiers: GMM, K-NN and BT.

The usage of less classifiers and the discarding of some features reduces the amount of tests significantly but they are still very numerous. It is calculated that, during this step around 272.276 unique tests were carried out per classification. As three different classifications make up this step, the total number of tests ascends to 816.828. Even with less classifiers and feature parameters to test, the amount of tests is really big due to the three different classifications, this makes the usage of powerful servers essential.

5.7.2 Results

The results section is divided into the three different classifications devised earlier. It should be noted that the same approach was followed in terms of graphs as in the previous classification: only the top five for each type are shown with the intention of reducing the clutter.

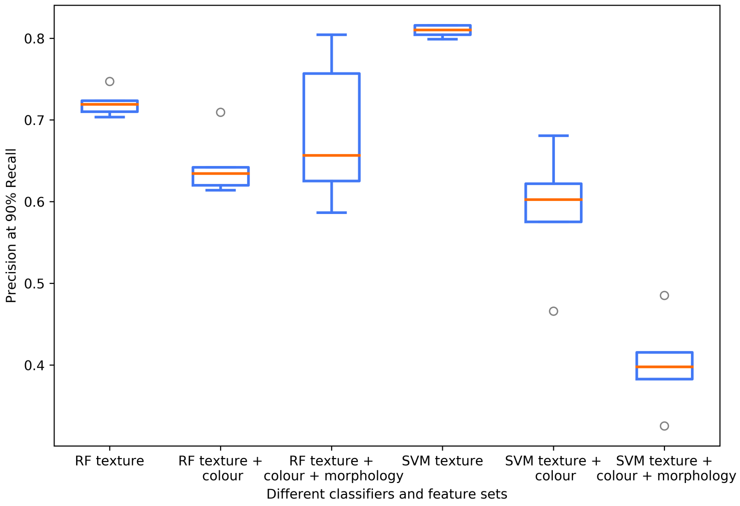


Figure 5.13: Results when classifying *W. naegeliana*

Results of W. naegeliana versus the rest

In figure 5.13 the results for each classifier with the different set of features are presented. It can be seen that both RF and SVM are really close, with the SVM with only texture features getting the overall best result with 82% of precision at 90% of recall.

In this step the performance of the unmasked approach was worse than the masked as they reached a maximum precision of 69% and 81% at 90% recall, respectively in the SVM classifier. In the random forest the result is similar with 67% with the unmasked images while masking them reported 74% of precision at 90% recall. A graph detailing the results of both approaches can be found in figure B.2, located in the second appendix.

Results of harmful vs harmless organisms

In figure 5.14 the results of this classification are presented. In that figure the performance of the different classifiers with the different feature sets can be seen. The top performer is SVM with just texture features with a 83% of precision at 90% of recall.

In this classification the difference between the unmasked and the masked versions is really negligible. The highest performance by an unmasked system is 81% while masking the image yields a the 83% previously mentioned. More examples of the performance difference between both feature sets can be found in figure B.4, located in the second appendix.

Results of Dinobryon Sp. versus the rest

The results of this classification are presented in figure 5.15. Many different feature sets obtain similar results but the highest performance is obtained with the random forest classifier using

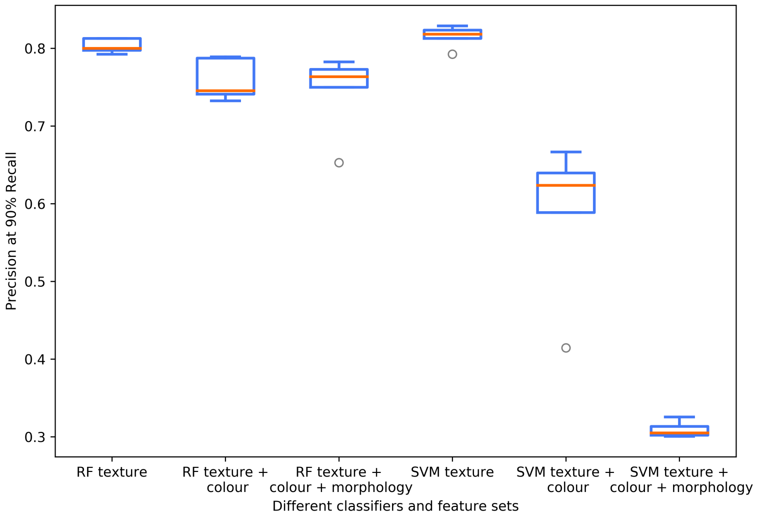


Figure 5.14: Comparison of RF and SVM with different features when classifying phytoplankton in harmful and harmless groups

all three features, texture, colour and morphology, an 80% of precision at 90% of the recall. The results for RF with texture only and SVM with colour and texture are really close to it, both being around 79% of precision at 90% of recall. The masked and unmasked approaches are also really even as both obtain around the same performance, 79% for both when using only texture. Extended information about the performance of masked and unmasked features can be seen in figure B.3, in the second appendix.

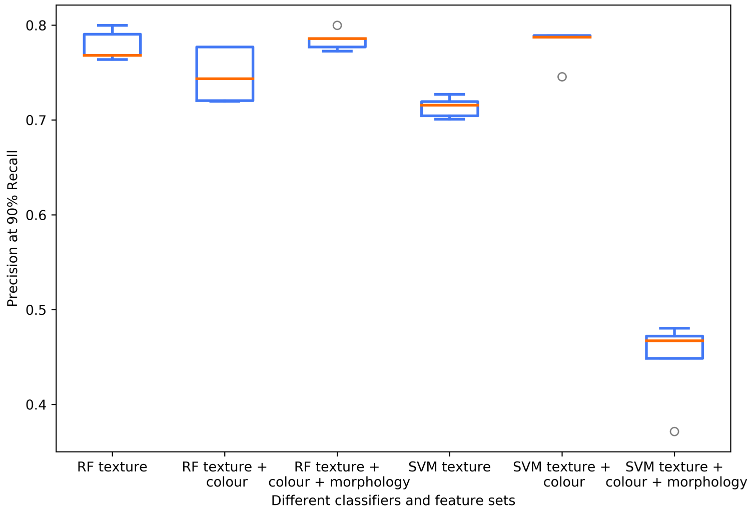


Figure 5.15: Results with different classifiers and features when classifying *Dinobryon Sp.*

5.7.3 Discussion

Overall, the results for the three different classifications are all satisfactory. The highest precision scores are obtained by the *W. naegeliana* and the harmful classification, both being very similar. The *Dinobryon Sp.* classification is not far behind with a 79% of precision when compared to the 82% of both previous groupings.

The results for all the classifications show higher precision than the ones from previous step. As hypothesised previously, it is likely due to the heterogeneity of the phytoplankton as a whole. This problem persists, to a lesser degree on this step but the results are still good despite the obstacles.

In this step, the features other than texture showed their usefulness with some of them scoring the highest precision like in the case of *Dinobryon Sp.* These features also accentuated the differences between both classifiers. With only texture features both yielded similar results but paring SVM with the extra features never did report benefits while RF sometimes did benefit from said extra features. This can be appreciated in the figures 5.14, 5.13 and 5.15, where the boxplots for each classifier and feature set are laid out.

This set of classifications also validated the usage of masked features. While in the previous phytoplankton vs non-phytoplankton classification the masked versions always scored lower than their unmasked counterparts in this step the masked features offer higher precision for bot the*W. naegeliana* and the harmful classification. In the *Dinobryon Sp.* both feature sets are more or less on par. This is most likely due to the lower quality of the masks for *Dinobryon Sp.* organisms as they are generally transparent and thus more difficult to segment than other phytoplankton.

5.7.4 Summary and conclusions

In this section the proposed methodology to classify the specimens into different groups managed to, again, obtain good results despite the difficulty of the problem. Each classification had its complication and purpose but they were successfully tackled.

The main takeaway from this step is that all the features showed good performance in some instance, showing their validity and usefulness. Both classifiers showed similar performance with some changes as SVM did not perform very well with colour or morphological features while RF often improved.

Overall the results obtained in this section are quite satisfactory and good in the three classifications tested. Each one of them has their particular difficulties that were overcame with the different combinations of features and classifiers. Even if the problem is complicated due to the variation of specimens in the same species group the performance was very good given all the features that this project has, like the overlapping between multiple elements.

Despite everything, each classification got a high precision, with harmless vs harmful and *W. naegeliana* topping at around 82% and *Dinobryon Sp.* at 79%, both at 90% of precision.

5.8 Summary and conclusions

In this chapter the methods for classifying the detections of the chapter 4 was studied. Several features were extracted from the images which include:

- **Bag of Visual Words with Gabor filters:** texture features
- **Bag of Visual Words with RGB channels information:** colour features
- **Number and proportion of pixels in the mask:** morphology features

Several classifiers were also tested with these features. The classifications were divided into two groups: phytoplankton detection and species classification. Phytoplankton detection aims to discern phytoplankton from other detections like zooplankton and garbage. Species classification intends on distinguishing species of phytoplankton to perform potability testing. Species classification holds three different classification, *W. naegeliana* versus the rest, *Dinobryon Sp.* versus the rest and harmful versus harmless organisms. Each one of those groupings has their particular characteristics that pose a particular obstacle to obtain the correct classification.

Both phytoplankton detection and species classification show good performance in the chosen metric, precision at a high level of recall, like 90%. This metric intends to lower the false negative rate in order to preserve a high amount of dangerous specimens, the base for potability testing.

Some of the main challenges of this work are the massive inner-class variability while having to deal with inter-class similarity. The difficulty of segmentation due to overlapping between objects can also impact the results. Many tests were made with alternate features and classifiers, nearing a million and a half, in order to find the best combination of parameters. Despite the complications, the results for every classification were good, a 77% for phytoplankton detection, a 79% for *Dinobryon Sp.* and around 82% for both *Woronichinia naegeliana* and harmless vs harmful. All this previous metrics are precision at 90% of recall.

The performance of the system is good and overall the difficulties can be overcame but more data may be needed to pursue classification of other species and to perfect the current ones.

Chapter 6

Conclusions

The objective of this project is to to create an innovative method to process digital microscopy images of phytoplankton organisms. The focus is to provide a suitable system to work in a real world scenario, and provide an appropriate evaluation for that purpose. The idea is to create a system that can perform a significant part of phytoplankton analysis automatically. The main target is the potability tests for water. Water may become toxic due to the presence cyanobacteria that need to be located and counted to determine their concentration per litre of water and thus the water quality. Any kind of analyses requires a lot of time and knowledge from experts, that due to the lack of a completely exhaustive and specified method, tend to be inconsistent. This may be specially important in potability testing since it should be carried out quite frequently to monitor water reservoirs and public health is on the line. The toxins that this phytoplankton produce can cause problems for humans, from mild ones like gastroenteritis to very severe ones like cancer or just death. It is clearly important then that this analyses are done regularly and without any bias. This project aimed to create a system to aid in those tasks while experimenting with new techniques and open a line of research into the relatively ignored world of freshwater phytoplankton segmentation and classification.

The whole process started with a broad domain analysis ranging from better knowing what were the problems caused by the cyanotoxins and the fact they could not be eliminated from water with normal water purifying techniques. Later a thorough state of the art analysis revealed that even-though there are many works related to plankton analysis there is a lack of works in the particular scope of freshwater phytoplankton. But more relevant is the fact that most of the existent works do not allow to use conventional microscope imagery. Many of the current methods instead, use several types of special hardware which allow to gather samples of water, take pictures of phytoplankton while in-flow or segment the phytoplankton specimens on the water trough images. Therefore, the main focus and methodology of this work is completely innovative as there is no single work that aims to perform a similar task using this type of images.

While the main focus of the work is the potability of the water it also opens a whole new line of research into the classification of species. The advantage of this tool would be its easiness of use as it does not need for any special hardware or big deviations from the current practises. The experts would normally sample water and capture the images but instead of manually identifying microorganisms and counting them, they would just pass the image to the software. To carry out this project the set of microscopy images was taken from real-world samples and under real imaging conditions. This lake is subject to frequent potability tests, so this type of system could prove useful. The use of real world samples meant that the results of the experiments would be a real case, not an ideal one. This marks a big difference with many state of the art works that use curated or ideal dataset often focusing on too much of a niche.

One of the main sources of complication for this project is related to the fact that the images are real ones as they are often filled with garbage and all types of undesired elements. Other source of complication in this project is the fact that these images feature many specimens randomly distributed across each image. This, in itself, represents an innovation as it is not common in the state of the art for a work to take this approach. All the elements on the images and their positions relative to one another increase the difficulty of the segmentation since many of the specimens overlap to each other which makes it nearly impossible to obtain a correct segmentation. It is also important the fact that phytoplankton are deformable and their shape is subject to multiple changes. Many of the features of phytoplankton are also shared with zooplankton, also present in the images and even some of the more common garbage types can look very similar to phytoplankton specimens. All of these are the reasons that this type of analysis needs a well trained expert and a lot of experience to produce correct results.

The project was divided into two interrelated parts, first the segmentation of the image that aims to distinguish every specimen from the background, and then the classification of the detected specimens. This classification is itself divided in two parts, the first one aims to clean the set of detected specimens separating the phytoplankton ones from the rest. The second one focuses on species classification with special focus on the dangerous ones.

The segmentation proved complicated but it was solved with good results. The designed process is able to accurately detect specimens and fuse them when needed into the so called colonies. Several methods were tested in order to take on the complicated nature of the images as they contain many and varied specimens. It was necessary to use an adaptive threshold method as simpler methods failed due to irregularities on the image illumination. Otsu's method was also tested and discarded as images showed variations in the levels of foreground-background proportion which can not be controlled as it entirely depends on the water samples. The adaptive threshold algorithm makes use of all the RGB channels in order to gather more information and obtain more reliable detections. After this, it was necessary to create an specific algorithm to deal with the colony fusion as many of the specimens group up in said colonies. The algorithm is based on the similarity between candidates using colour and distance. The overall results are good with the system only missing a few specimen representing a 0.256% of false positive rate. This comes at the cost of detecting many false positives which are cleared out in the classification step. The quality of the detections is also good with limited cases of over and undersegmentation, totalling 3.42% and 12.53% respectively. Overall the performance of the system is good despite the difficulty of the problem as the many specimens per image and their relatives positions (often overlapping) increase the difficulty of the problem.

The second part, classification, aims to clean the detections and classify them into their species. To accomplish this several features were experimented with as well as different types of classifiers. Many combinations of these were tested with SVM and RF resulting to be the best classifiers and texture features proving to be the most important ones. Colour features also proved to be useful although not as important as the texture ones. Both of these features are based around the Bag of Visual Words model. As the amount of dangerous or interesting specimens lost is the most important metric the precision was evaluated in relation to the recall. This part particularly proved to be of great complication due to the high variation in both the phytoplankton and garbage groups while some types of them were very close, as explained before. Despite the complication of this step the performance was very good with the top result being a 77% of precision at 90% recall.The second part focusing on species separating *W. naegeliana* and *Dinobryon Sp.* from the rest,a third classification separates harmful and harmless groups. Across the board, all the classifications obtained good results thus demonstrating the validity of the features used. While *W. naegeliana* versus de the rest and harmful vs harmless managed to obtain around 82% of precision at 90% recall, *Dinobryon Sp.* versus the rest got a 79% of precision at 90% recall. All of this results highlight the validity of the system despite the high amount of complication that this problem exhibited.

All in all, this project tackled a important currently unsolved problem such as phytoplankton segmentation and classification from a real world point of view with all the handicaps that it adds and reached a satisfactory solution. Innovating in the methods and the data that the project used as the state of the art lacked an approach that could use conventional microscopy images to detect and classify phytoplankton specimens. This work is a great step to create a non specific phytoplankton segmenter and classifier. Both steps of the work, even if not flawless, showed great promise and performance when trying to solve a very complex problem. The system is robust, a essential part for a system that aims for real world deployment. While this work presents a complete, beginning to end, tool which has all of its components in a working state with good experimental results in each and every one. This should allow to reduce part of the work that experts need to do without some of the drawbacks that other automatic systems have, like price of bulk. This work serves as an initial development in this line of research even if further improvements are still possible. All of this will be detailed in the chapter 7, but as conclusion to this work, it represents a good foundation to continue development and research in this topic.

Chapter 7
Future work

This work serves as foundation for new ideas and possibilities in this research line. The easier one to see is related with the amount of data used in this work. It becomes apparent in the second classification section 5.7.2 that the images where not enough to create a full classification. The first step should then be to get more images for the dataset. The used images are all from the same sample of water, so to increase the amount of images and variability in the organisms present samples from across the year will be desirable. It is also important to use samples from different reservoirs or lakes to not limit the scope too much. While this work has focused on toxic species present in Galicia, it could possibly be extended to other territories. In the end, one of the most important components of this type of works is good quality data representative of a real world scenario.

In terms of the first step a good starting point would be to get a methodology to clean up all images. This is relevant due to the fact that the changes in illumination make some techniques significantly worse. Another model of microscope might produce other artefacts not accounted for in this work. The current methods are very robust so they should hold up even with different lighting settings or other type of noise, but to test it new data is required.

In the segmentation step, many methods were tested however, as mentioned in chapter 4 some more features or data could be added to the graph pruning algorithm. Currently it only measures similarity through distance and colour but other features may improve its already good performance, like shape or other morphology descriptors.

On the classification step, other methods can be tested. During this work many classifiers were employed like RF and SVM. Artificial Neural Networks (ANN) were not tested due to a time concern with its training. Theoretically, the tested classifiers same capabilities of solving this types of problems as ANN, however testing this method is still pending and could report some benefits.

Testing deep learning methods is also still pending as simpler methods were tested before in order to asses the complication of the problem. After the fact, it is clear that the problem analysed in this dissertation is far from simple or trivial but work needed to be done before jumping to conclusions, so deep learning could prove beneficial.

As for the features, many of them were thoroughly tested but adding others may improve the system. For texture, local binary patterns could be useful. Other, more complicated, morphology features may also be of use like convexity, circularity or even devise a new one for the particular organisms that need to be classified. The latter one is often used in the state of the art articles. Colour is another feature where some more experimentation could benefit results. Currently a bag of words with the three RGB channels is used but maybe some further experimentation using other descriptors like taking the colour of certain relevant points using something like SIFT or simply computing the dominant colour of different regions. With the new features a feature selector could be added to simplify what features are used by the classifier. Another possibility is to use a feature selector which chooses the most representative features from a large feature vector.

In conclusion, this work serves a starting point for new research, analysing a topic still relatively unpopulated, so many new ways of continuing this analysis arise. The scope of this project was not enough to test everything out and many of these details are left for future works.

Appendices

Appendix A

Theoretical explanations

This chapter contains the theoretical explanations of some of the methods used in this project with more depth and detail than the more brief explanation found in the dissertation.

A.1 Otsu's thresholding method

Otsu's method is a global thresholding algorithm. It is able to automatically select a threshold value based on the image's histogram contents. This threshold value intends to minimise the intraclass variance or maximise the interclass variance, as they are both equivalent. Both of these metrics are weighted according to the size of each region

$$\sigma_w^2(t) = \omega_0(t)\sigma_0^2(t) + \omega_1(t)\sigma_1^2(t) \tag{A.1}$$

Where weights ω_0 and ω_1 are the probabilities of the two classes separated by a threshold t, and σ_0^2 and σ_1^2 are variances of these two classes.

The simplest version of this algorithm follows a series of steps that start with the creation of the histogram. The histogram is a probability distribution of the colours that make up the image. In the case of a grayscale image each bin of the histogram represents a different shade of Grey an its amount on the image. The weights seen in the equation A.1 are calculated in the histogram simply summing all the bins that corresponds to each one of the classes. The algorithm would iteratively step through all the possible thresholds computing the weighted variance as seen in A.1. The variance for each class and weighted according to the previous description. Once all the thresholds have been tested the one that produces the lower intraclass variance for both classes, minimising the result of equation A.1 is selected. Faster versions of the algorithm have been created allowing to reduce the exploration, time cutting down on the necessary variance tests. An example of the chosen threshold value and

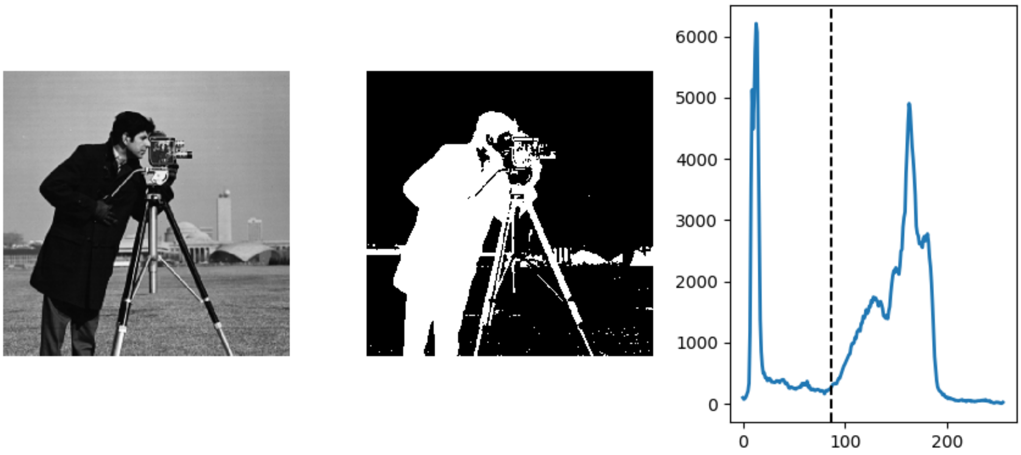


Figure A.1: Demonstration of Otsu's thresholding method

its representation on a histogram can be seen in figure A.1.

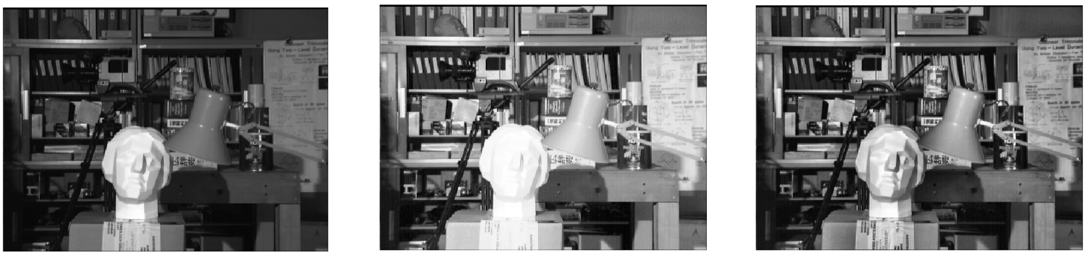
As mentioned in the dissertation, Otsu's method only performs well when images have a clearly bimodal histogram. This limits the usefulness of this, and other global methods, as it requires a similar amount of background and foreground.

Several improvements and versions of this algorithm have been made like Gaussian Otsu's method that is faster than the optimal Otsu's method [72].

A.2 Histogram equalisation and CLAHE

Histogram equalisation looks to improve the contrast of an image by evening out all the graylevel probabilities of the image's histogram. This means spreading out the most frequent intensity.

The simpler methods look to simply even out the histogram of the image. This is done by *stretching* the bigger intensity probabilities in a way that unpopulated or scarcely populated areas of the histogram receive more values. This allows, that, with the same information, more contrast is perceived as the differences between pixels have been made bigger. More advanced, and generally better techniques have already been developed, like Adaptive Histogram Equalisation (AHE). This technique is different from the global methods, like the one explained above, as it does its job in small parts of the image, also known as windows. These windows are small sections of the image that are picked subsequently, in order to cover each pixel of the original image. This way, each window has a particular transformation done, allowing for better performance as the changes are only based on that part of the image. Contrast Limited AHE (CLAHE) intends to solve a persistent problem with AHE systems, the over-amplification of the contrast in near-constant regions, which causes noise. CLAHE



(a) Original image (b) Global equalisation (c) CLAHE

Figure A.2: Histogram equalisation methods compared

solves this problem by clipping the histogram at a predefined value, often called clip limit. The clipped part of the histogram is not discarded as it is redistributed equally among all the other histogram bins. A comparison between these methods can be seen in figure A.2 where the improvement in the results can be easily seen.

A.3 Gaussian Adaptive thresholding

Adaptive thresholding works by taking small portions of the image, known as windows, and thresholding them separately. This algorithm is considered local as it only takes information from each window to perform the transformation on said part of the image, it does not take into account the rest of the image. This kind of algorithm is more sophisticated than global thresholding methods, as it can accommodate to the changing lighting conditions in the image, one of the reason why it has been used. This is due to the assumption that smaller image regions are more likely to have approximately uniform illumination.

This algorithm has several variations like the Gaussian adaptive threshold, the one used in this work. All the adaptive algorithms use a metric to select the threshold value, like the mean. In the case of the Gaussian version, it uses the weighted sum of neighbourhood values, with weights being a Gaussian function. The Gaussian version of the algorithm probed to be best in this work as it managed to collect far less noise than other metrics like the previously mentioned mean. This can be seen in figure A.3 where both methods are compared side by side.

A.4 RGB

A colour model is a visualization that represents the colour spectrum like a n-dimensional model, usually of three dimensions. RGB is an additive colour space defined by the RGB colour model which in turn is defined by three chromaticities of the red, green, and blue [73].

(a) Gaussian Adaptive Thresholding

(b) Mean Adaptive Thresholding

Figure A.3: Adaptive Thresholding methods compared

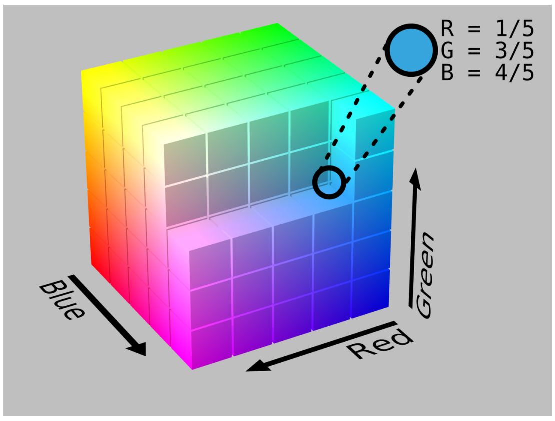


Figure A.4: Visual representation of RGB colour model as a cube

This means that this colour model is three dimensional. A representation of this colour model can be seen in figure A.4 [74].

In this model a colour image can be split into three different channels, each one of them containing information of a particular colour. The complex colours of an image are formed by mixing the information of each base colour contained in each channel. The information is mixed in an additive way, this means, that for example setting every colour to its maximum value will yield the brightest colour, white. Setting every colour to 0 results in black. An example of colour separation and its mixing can be seen in figure A.5 [75] which depicts a real image separated in its channels.

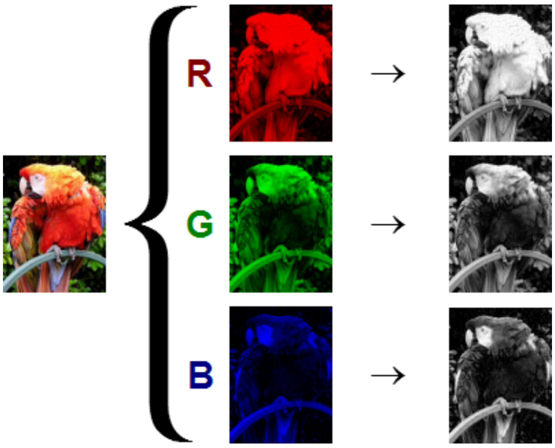


Figure A.5: Image separated in its RGB channels

A.5 Mathematical morphology

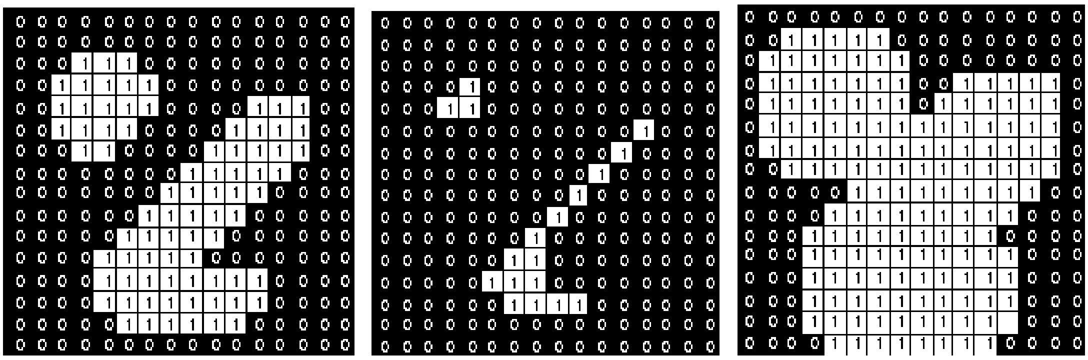
Mathematical morphology operators are based on set theory, these operators are widely used in the field of computer vision and image processing. Mathematical morphology operators are based off two basic operators after which the rest of them are modelled [76]. This methods are often used over binary images, like the ones used in this work, although they can also be used in grayscale images with slight modifications.

Mathematical morphology methods require a kernel or structuring element. Kernels are applied to the image treating both, the image and the kernel, as sets. Those sets are similar although different, for example, the kernel is normally smaller than the image it is applied to or it may have a wider range of values like negative ones when compared to the image, usually binary. It is also common that the kernel has its origin of coordinates in the middle rather than in a corner like a normal image. The centre is really important is it greatly influences the output of the different operators.

The base operators are erosion and dilation. Erosion deletes regions where the kernel *does not fit*, that is, regions smaller than the kernel are erased, effectively slimming figure. Dilation is the opposite of erosion as it enlarges the boundaries of the figures making them bigger. A comparison of both can be seen in the example provided in figure A.6. With both these basic operators other more complex ones can be created like opening, which is an erosion followed by a dilation, and closing, which is a dilation followed by an erosion.

A.6 Suzuki and Abe's algorithm

This algorithm is used to find contours in binary images. It is able to differentiate the outer edges of each contour and omit those inside them which made it ideal to use in this work as



```
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 1 1 1 0 0 0 0 0 0 0 0 0 0
0 0 1 1 1 1 1 0 0 0 0 0 0 0 0 0
0 0 1 1 1 1 1 0 0 0 0 1 1 1 0 0
0 0 1 1 1 1 0 0 0 0 1 1 1 1 0 0
0 0 0 1 1 0 0 0 0 1 1 1 1 1 0 0
0 0 0 0 0 0 0 0 1 1 1 1 1 0 0 0
0 0 0 0 0 0 0 1 1 1 1 1 0 0 0 0
0 0 0 0 0 0 1 1 1 1 1 0 0 0 0 0
0 0 0 0 0 1 1 1 1 1 0 0 0 0 0 0
0 0 0 0 1 1 1 1 1 0 0 0 0 0 0 0
0 0 0 0 1 1 1 1 1 1 1 1 0 0 0 0
0 0 0 0 1 1 1 1 1 1 1 1 0 0 0 0
0 0 0 0 0 1 1 1 1 1 1 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
```

(a) Original image

```
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0
0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0
0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0
0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0
0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0
0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0
0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0
0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0
0 0 0 0 0 1 1 1 0 0 0 0 0 0 0 0
0 0 0 0 0 0 1 1 1 1 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
```

(b) Erosion

```
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 1 1 1 1 1 0 0 0 0 0 0 0 0 0
0 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0
0 1 1 1 1 1 1 1 0 0 1 1 1 1 1 0
0 1 1 1 1 1 1 1 0 1 1 1 1 1 1 0
0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0
0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0
0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 0
0 0 0 0 0 1 1 1 1 1 1 1 1 1 0 0
0 0 0 0 1 1 1 1 1 1 1 1 1 0 0 0
0 0 0 1 1 1 1 1 1 1 1 1 0 0 0 0
0 0 0 1 1 1 1 1 1 1 1 1 1 0 0 0
0 0 0 1 1 1 1 1 1 1 1 1 1 0 0 0
0 0 0 1 1 1 1 1 1 1 1 1 1 0 0 0
0 0 0 1 1 1 1 1 1 1 1 1   0 0 0
0 0 0 0 1 1 1 1 1 1 1 1 0 0 0 0
```

(c) Dilation

Figure A.6: Basic mathematical morphology operators

many specimens have holes in them or are partly transparent.

The algorithm works by first detecting all the candidate edges over the binary image. Then, to differentiate its type, that is, if they are outer edges or inner edges it looks at their relation with other edges. The outer-most edge is the frame of the picture. This edge is parent of every other one in the image as encloses them. If any other random edge only has the frame as parent then it is an outer edge. However, if any other border has more than that one parent it means that it is a hole or inner edge and it is not accounted for. An explanatory image can be found in figure A.7.

This way, Suzuki and Abe's algorithm is able to detect only the phytoplankton specimens ignoring their holes which, in turn, makes it easier to create binary masks that will serve to detect the specimens after some post-processing.

A.7 Delaunay Triangulation

A triangulation is a process where a plane is divided into triangles following a certain criteria. This process can be extended to more dimensions as well, however the version used in this project operates over a plane. Any triangulation can be represented as a graph, that is, a collection of vertices and edges with their corresponding relations.

The Delaunay Triangulation is a method, that given a set of points, is able to create a triangulation where no point or vertex is inside any circumscribed circumference of any other triangle. The circumcircle of a triangle is the unique circle passing through the three vertices of the triangle. Triangles are, of course, made out of the points passed to the algorithm. The aim of this process is to subdivide the plane in triangles.

This triangulation is related to the Voronoi graph, as the vertexes in the regions of the former are the circumcentres of the Delaunay triangles. In this work the set of points used

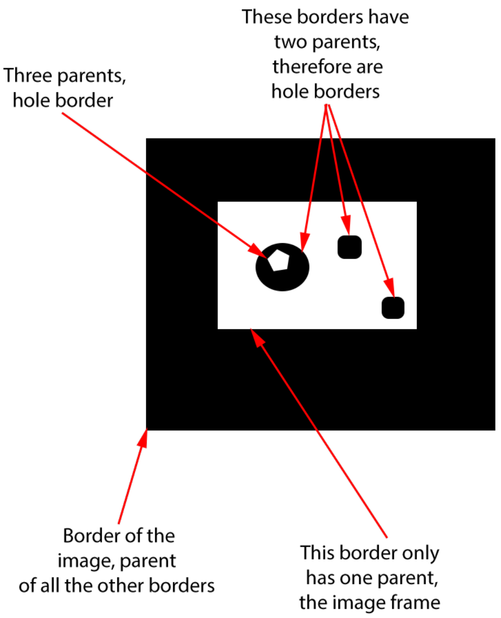

These borders have two parents, therefore are hole borders
Three parents, hole border
Border of the image, parent of all the other borders
This border only has one parent, the image frame

Figure A.7: Demonstration of the Suzuki and Abe's algorithm

to create this graph is the centroids of the detected contours. The centroid of a figure is the point where all the planes that pass through it divide the figure into two equal parts. Using the centroid allows to get a representative point of a figure in any shape, which is useful due to the varied shapes that these detections can take.

While a concrete example of this triangulation for this project can be seen in figure 4.15 a more general example that features the representation of the circumcircles is shown in figure A.8 [77]

A.8 Bag of Words model

The Bag of Words model originated in the field of information retrieval and natural language processing as the means of simplifying the representation of a text. With this model a text is broken down into its desired components like words or phrases and transformed into a bag. This bag only keeps track of how many different instances of each appear. The bag acts like a vocabulary or dictionary of the text. This bag can then be used to extract features from a text that in turn can be used for modelling. An example of how this model works can be seen in A.8.1

As explained in the dissertation the basic idea of this model has been adapted into the field of computer vision through the Bag of Visual Words, which is used in this work. This uses the same concepts but instead of words uses features from the images which can be extracted through different algorithms.

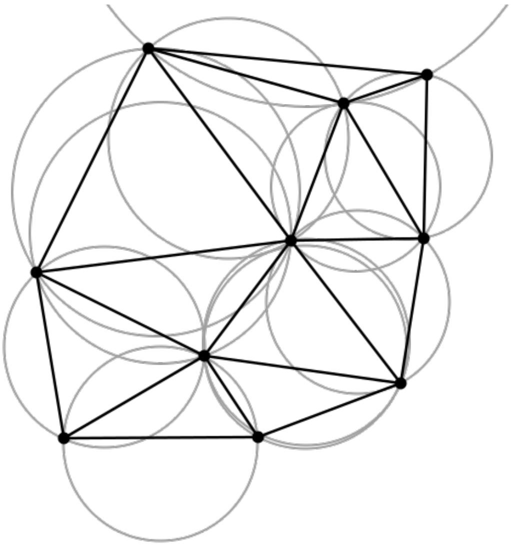


Figure A.8: Delaunay triangulation showing no vertex inside any circumscribed circumference

A.8.1 Example of the Bag of Words model

Phytoplankton are part of the plankton.
Phytoplankton are the foundation of the oceanic food chain.
Cyanobacteria are a type of phytoplankton.

The unique words forming this sentences are: *phytoplankton, are, part, of, the, plankton, foundation, oceanic, food, chain, cyanobacteria, a, type*

The result of transforming each sentence into a vector of its words yields:
Phytoplankton are part of the plankton = [1,1,1,1,1,1,0,0,0,0,0,0,0]
Phytoplankton are the foundation of the oceanic food chain. = [1,1,0,1,2,0,1,1,1,1,0,0,0]
Cyanobacteria are a type of phytoplankton. [1,1,0,1,0,0,0,0,0,0,1,1,1]

A.9 Classifiers

All the classifiers used in this work are described in more detail below:

A.9.1 SVM

Support Vector Machine or SVM in short are a type of classifiers that are based upon the notion that the data points in an n-dimensional space can be separated by a hyperplane that

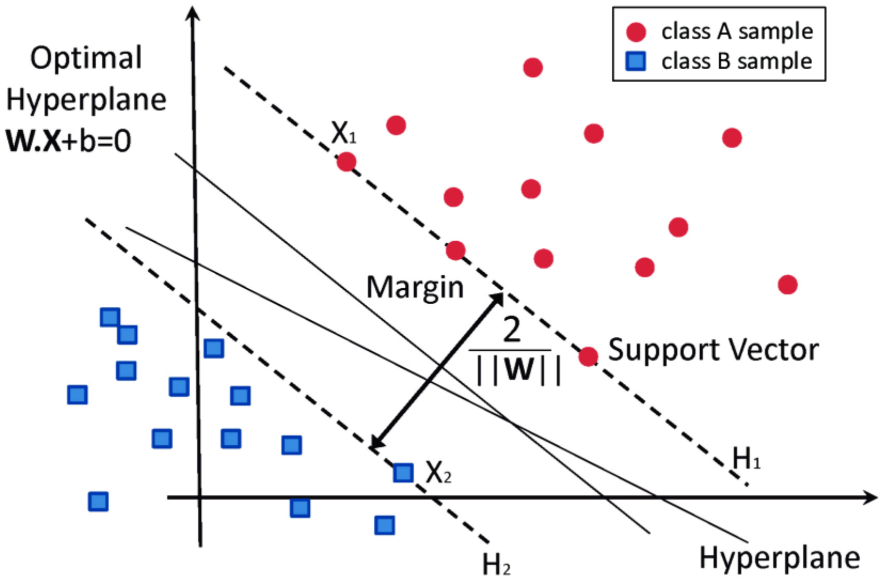


Figure A.9: SVM linearly separable example

maximises the space between said hyperplane and the closest data points [78]. These data points are called support vectors, and they determine the position and orientation of the hyperplane. SVM is characterised by the maximization the margin between the support vectors and the dividing hyperplane. An example of this can be seen in figure A.9 [79].

A kernel function allows to map the data points to a higher dimensional space. This could result in improved performance if in the current number of dimensions the classes boundaries are not linearly separable. An example of this can be seen in the figure A.10 [80] that shows a non linearly separable dataset that, with the appropriate kernel can be satisfactory split into the classes that make it up. This classifier type only needs to save the support vectors which makes them lighter to operate with than others classifiers like like k-NN.

A.9.2 kNN

K Nearest Neighbour classifier or k-NN [81, 82] is a classifier based on the idea of classifying data points through similarity. In order to classify a new sample, it takes the k nearest neighbours in the feature space, among a set of samples for which the class is known. The class is decided by voting, that is, if a sample has three neighbours of one class and two of other the majority class would win. The number k represents the number of neighbours so, to avoid ties, k is usually odd. Both the voting and the similarity function are used to asses the neighbours and may be varied depending on the target application. Some examples of common similarity metrics can be Euclidean, Manhattan, Minkowski, etc. The distance between samples is inversely proportional to the similarity, that is, the more similar the less distance.

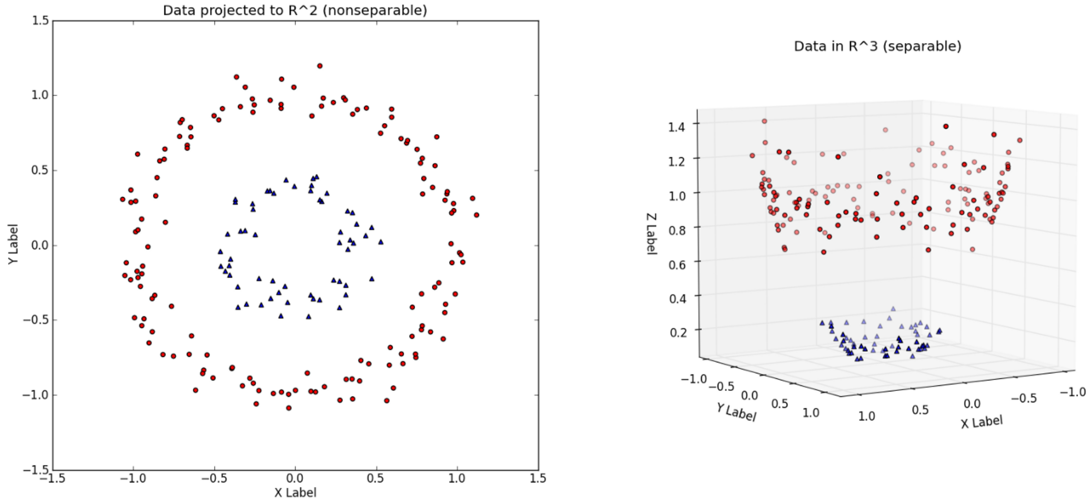


Figure A.10: An increase in dimensionality can make a dataset separable

An example of how this classifier works can be seen in figure A.11.

The number k also has a repercussion on the performance of the system [82]. If it is too low, the algorithm may take noise into account for the voting as it only looks few of its neighbours. Those neighbours could be outliers or just noise, which would make the classification erroneous. If the opposite happens and k is too big, it will bias the classification towards the largest class. This is due to the largest class having more examples which makes the decision skew towards it.

A.9.3 BT

Boosting trees (BT) [83] are an evolution of the earlier and more basic boosting algorithm [84]. Boosting is based on the simple affirmation that a series of weak classifiers can be united to create a stronger one [83]. While these weak classifiers have a higher probability of guessing correctly than the random guess, they are only loosely correlated with the correct classification. In the boosting process the samples are re-weighted and adjusted so that if they were misclassified using an early weak classifier, those samples gain higher weight in a later classifier. This allows the latter batch of classifiers to concentrate on previously wrongly classified examples. The key of this type of classifiers is that the weak learners are not made independently, like in a bagging system, but sequentially.

Boosting trees applies this principals to a series of decision trees. A decision tree is a structure shaped similarly to a flow-chart. It works by dividing the population into two or more populations that can be separated by a feature. An example of simple decision tree can be seen in A.12.

The differences between boosting and bagging algorithms can be seen in the visual exam-

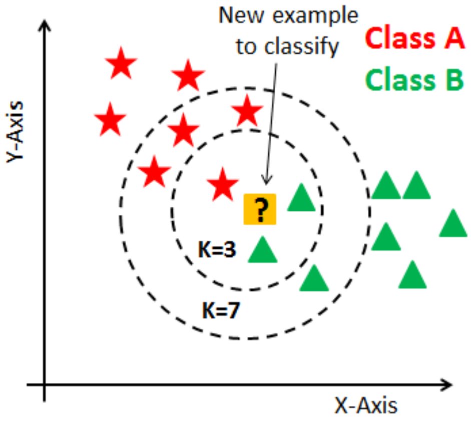


Figure A.11: Example of how a kNN classifier works

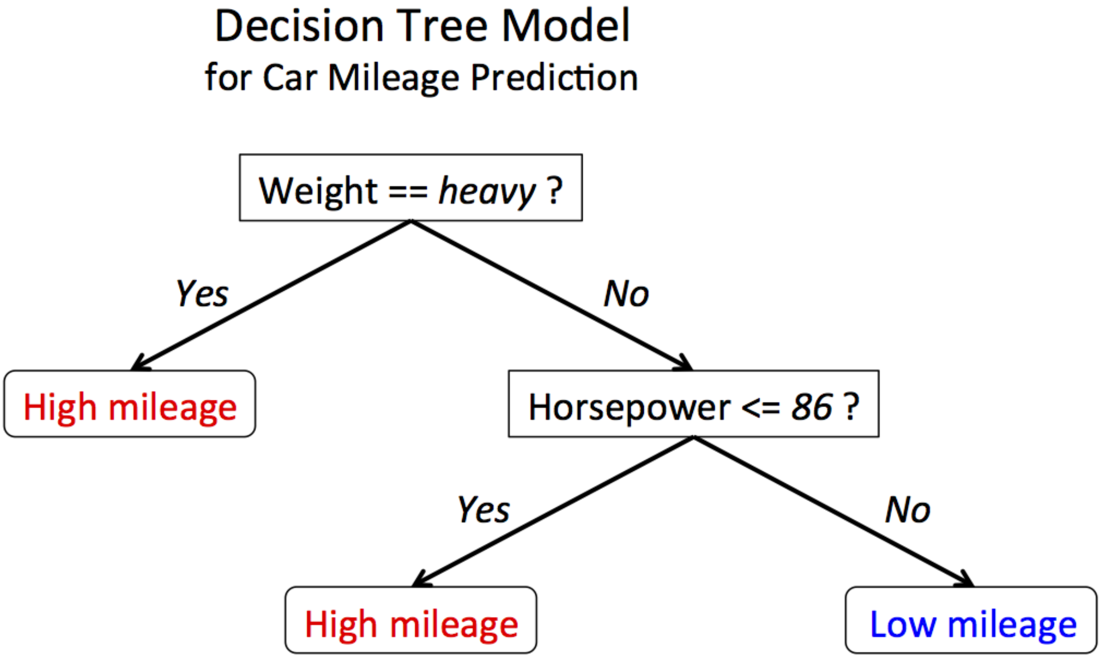


Figure A.12: Simple decision tree

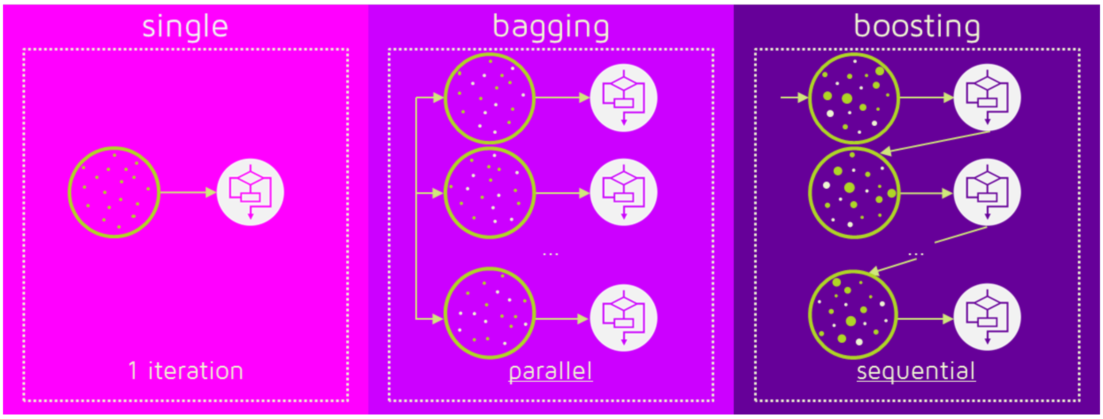


Figure A.13: Difference between boosting and bagging, basis for Boosting Trees and Random Forest respectively

ple contained in figure A.13 [85].

A.9.4 RF

Random Forest or RF for short is similar to Boosting Trees as it is also based around the idea of grouping learners. It groups, like boosting trees, decision trees that act as the weak classifiers in both cases. An example of simple decision tree can be seen in A.12. Random forest is not a boosting algorithm but rather a bagging one [86]. This means that the weak learners are not successively arranged. Instead, they are built independently and later combined trough a set of weights or averaging techniques like weighted average, majority vote, etc. Random forest [87] follows the bagging paradigm closely, only deviating in the fact that the tree learners in the learning process select a random subset of all the features which is called feature bagging. This process is done to find out the more correlated features that make up the stronger trees. The main disadvantage of this algorithm is that it tends to overfit to training data [84].

A visual example of the differences between boosting and bagging can be seen in the figure A.13 [85].

A.9.5 GMM

Gaussian Mixture Model or GMM for short [88] models the distribution of the data points assuming that they are generated from a mixture of a finite number of Gaussian distributions with unknown parameters. The adjustment process of a GMM to the data follows an Expectation-Maximisation (EM) [89] algorithm, similar to k-means, but it additionally optimises information about the co-variance structure of the data, apart from the centroids (the centres of the latent gaussians).

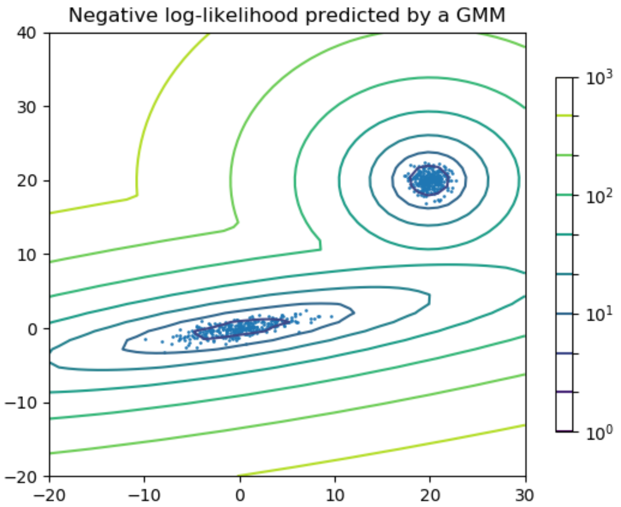


Figure A.14: Two-component Gaussian mixture model

This classifier uses the Bayes rule to determine to which of the groups a particular sample belongs to. This allows GMM classifiers to develop complex but smooth classification boundaries.

An example of this type of a Gaussian mixture model can be seen in figure A.14 [90].

Appendix B

Extra figures

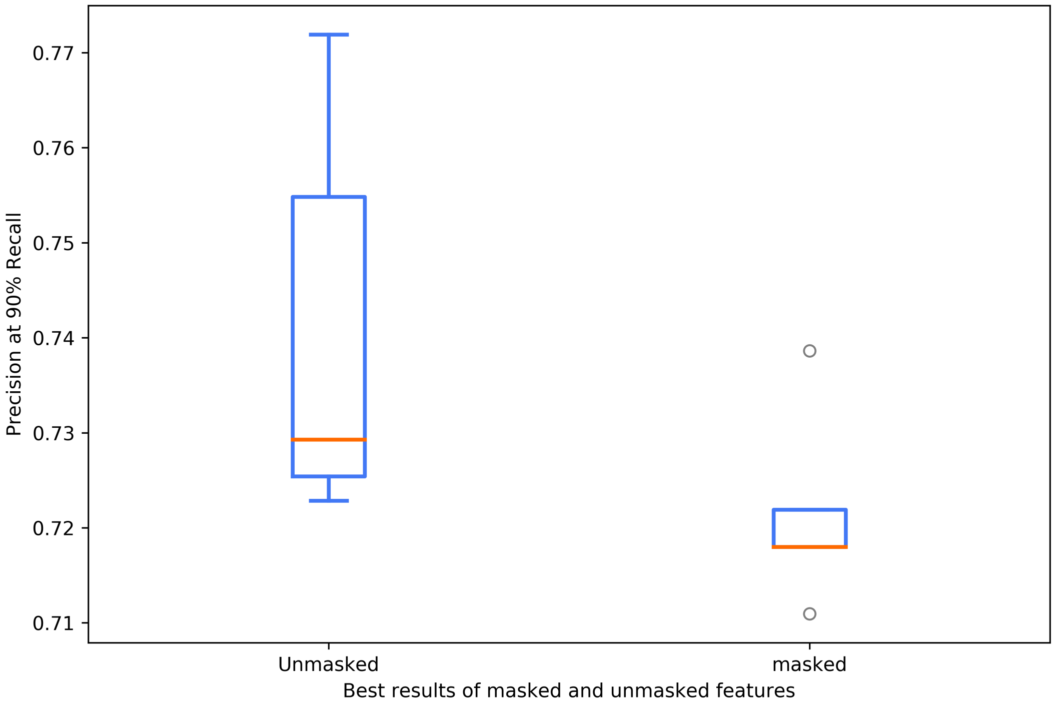


Figure B.1: Comparison of the five best results for unmasked and masked features for the first classification

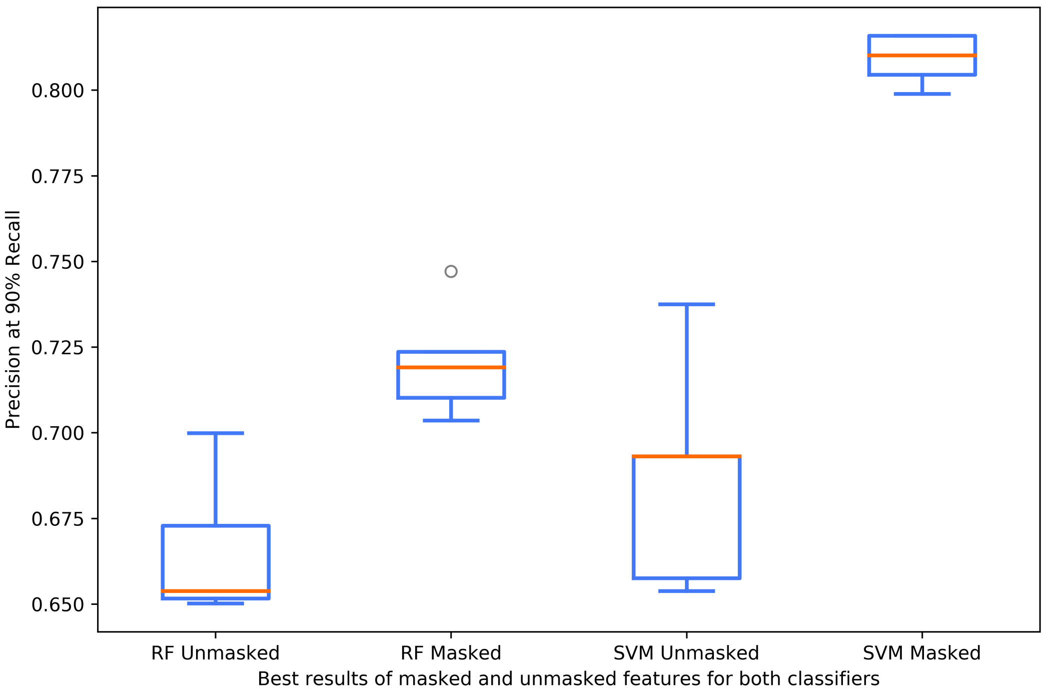


Figure B.2: Comparison of the five best results for unmasked and masked features and for both classifiers RF and SVM in the *W. naegeliana* classification

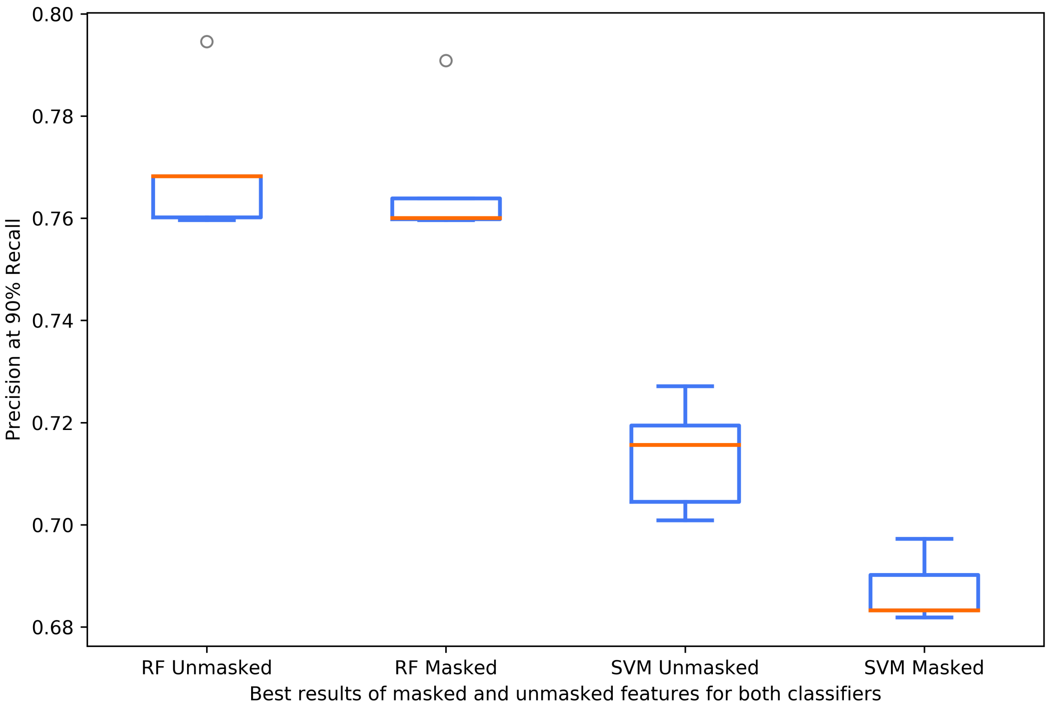


Figure B.3: Comparison of the five best results for unmasked and masked features and for both classifiers RF and SVM in the *Dinobryon Sp.* classification

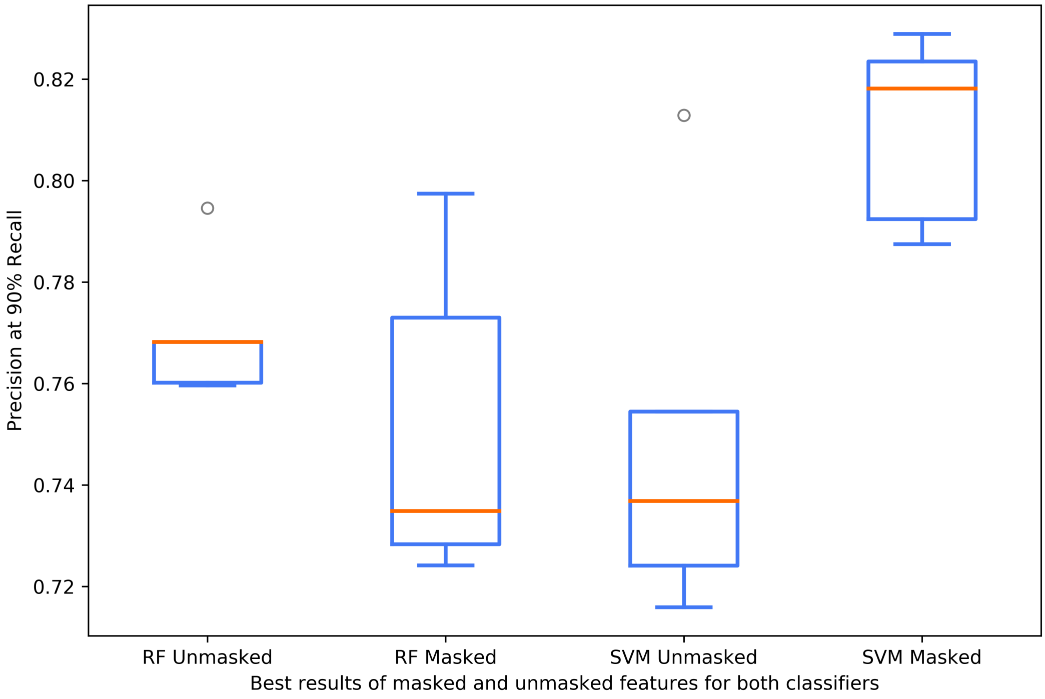


Figure B.4: Comparison of the five best results for unmasked and masked features and for both classifiers RF and SVM in the binary classification

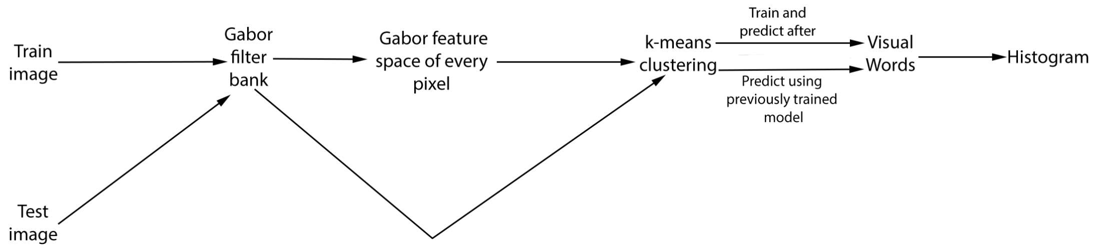


Figure B.5: Global process followed by the BoVW model

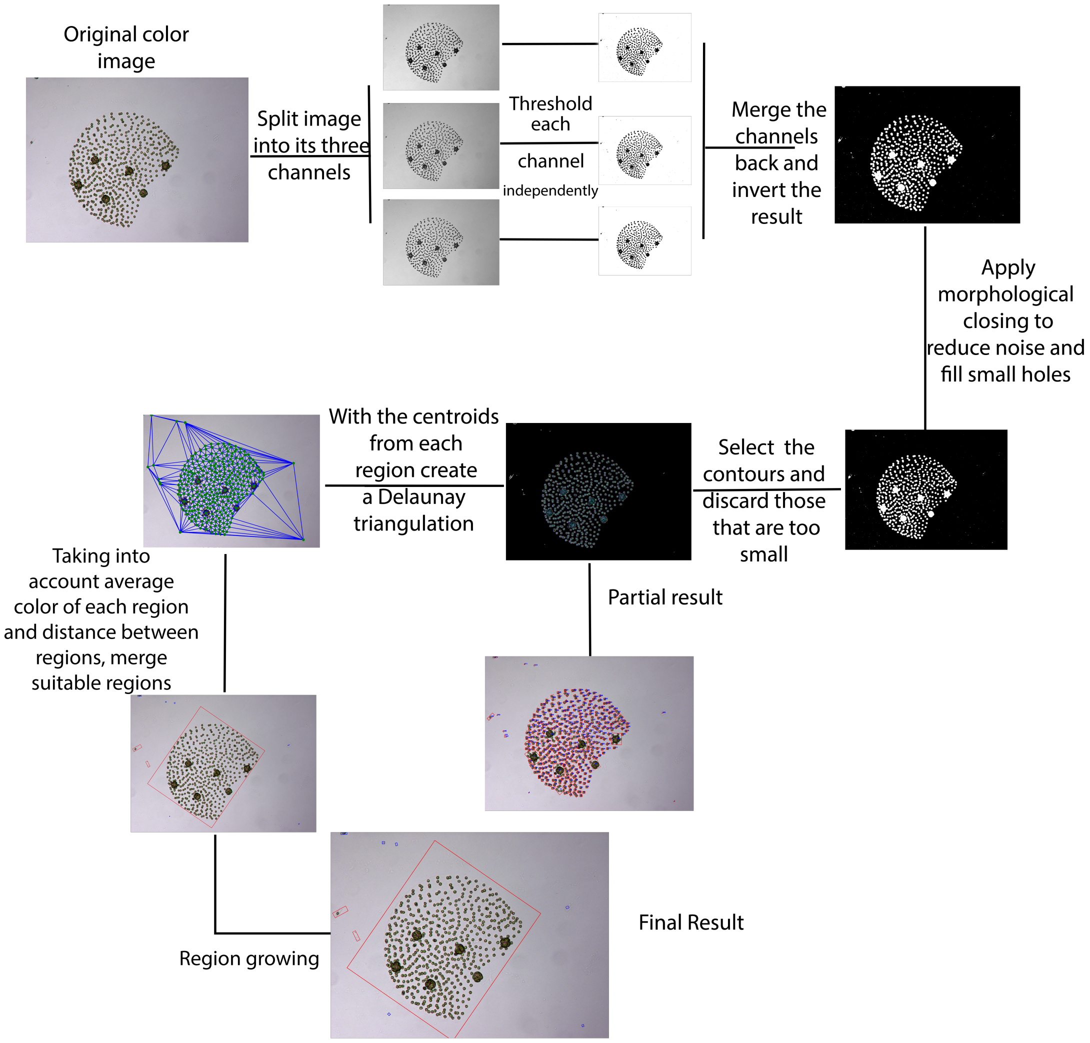

Original color image
Split image into its three channels
Threshold each channel independently
Merge the channels back and invert the result
Apply morphological closing to reduce noise and fill small holes
Select the contours and discard those that are too small
With the centroids from each region create a Delaunay triangulation
Partial result
Taking into account average color of each region and distance between regions, merge suitable regions
Region growing
Final Result

Figure B.6: Global vision of the segmentation process

Bibliography

[1] C. M. Lalli and T. R. Parsons, *Biological oceanography: an introduction*, 2nd ed., ser. Open University oceanography series. Amsterdam: Elsevier Butterworth-Heinemann, 2006.

[2] B. A. Whitton and M. Potts, Eds., *The Ecology of Cyanobacteria: Their Diversity in Time and Space*. Springer Netherlands, 2002. [Online]. Available: https://www.springer.com/gp/book/9780792347354

[3] J. Meriluoto, L. Spoof, and G. A. Codd, Eds., *Handbook of cyanobacterial monitoring and cyanotoxin analysis*. Chichester, West Sussex: Wiley, 2017.

[4] J. Roach, "Source of Half Earth's Oxygen Gets Little Credit," *National Geographic News*, Jun. 2004. [Online]. Available: https://news.nationalgeographic.com/news/2004/06/source-of-half-earth-s-oxygen-gets-little-credit/

[5] A. Zamyadi, F. Choo, G. Newcombe, R. Stuetz, and R. K. Henderson, "A review of monitoring technologies for real-time management of cyanobacteria: Recent advances and future direction," *TrAC Trends in Analytical Chemistry*, vol. 85, pp. 83–96, Dec. 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0165993615300406

[6] S. Cirés and A. Quesada, *Catálogo de cianobacterias planctónicas potencialmente tóxicas de las aguas continentales españolas*. Madrid: Ministerio Medio Ambiente y Medio Rural y Marino, 2011.

[7] L. Carvalho *et al.*, "Sustaining recreational quality of European lakes: minimizing the health risks from algal blooms through phosphorus control," *Journal of Applied Ecology*, vol. 50, no. 2, pp. 315–323, 2013. [Online]. Available: https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/1365-2664.12059

[8] C. de Hoyos, A. I. Negro, and J. J. Aldasoro, "Cyanobacteria distribution and abundance in the Spanish water reservoirs during thermal stratification," *Limnetica*, vol. 23,

pp. 119–132, 2004. [Online]. Available: https://www.limnetica.com/documentos/limnetica/limnetica-23-1-p-119.pdf

[9] I. Chorus and J. Bartram, Eds., *Toxic cyanobacteria in water: a guide to their public health consequences, monitoring, and management.* London ; New York: E & FN Spon, 1999.

[10] H. W. Paerl and V. J. Paul, "Climate change: Links to global expansion of harmful cyanobacteria," *Water Research*, vol. 46, no. 5, pp. 1349–1363, Apr. 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0043135411004386

[11] Ministry of the Presidency, "Directiva 2000/60/ce del parlamento europeo y del consejo, de 23 de octubre de 2000, por la que se establece un marco comunitario de actuación en el ámbito de la política de aguas." pp. 1–73, Dec. 2000. [Online]. Available: https://www.boe.es/buscar/doc.php?id=DOUE-L-2000-82524

[12] ——, "Directiva 2006/7/ce del parlamento europeo y del consejo, de 15 de febrero de 2006, relativa a la gestión de la calidad de las aguas de baño y por la que se deroga la directiva 76/160/cee." pp. 37–51, Mar. 2006. [Online]. Available: https://www.boe.es/buscar/doc.php?id=DOUE-L-2006-80413

[13] SERGAS, "Actuacións nos abastecementos de auga de consumo público que capten dun encoro con risco de crecemento de cianobacterias." [Online]. Available: https://www.sergas.es/Saude-publica/Documents/1520/ProtActAbastCiano.pdf

[14] Xunta de Galicia, "Plataforma Galega de información ambiental." [Online]. Available: https://gaia.xunta.es/plataforma/temas/agua/roaga/seguimiento-embalses/consulta-embalses

[15] A. Quesada, D. Sanchis, and D. Carrasco, "Cyanobacteria in Spanish reservoirs. How frequently are they toxic?" *Limnetica*, vol. 23, pp. 109–118, 2004. [Online]. Available: https://www.limnetica.com/documentos/limnetica/limnetica-23-1-p-109.pdf

[16] "La Xunta levanta la alerta por cianobacteria en el embalse del Umia activada desde julio," *Faro de Vigo*, Feb 2019. [Online]. Available: https://www.farodevigo.es/portada-pontevedra/2019/02/15/xunta-levanta-alerta-cianobacteria-embalse/2052140.html

[17] "El embalse de as forcadas está en alerta por proliferación de cianobacterias desde el 20 de mayo," *Diario de Ferrol*, Jul 2015. [Online]. Available: https://www.diariodeferrol.com/articulo/ferrol/edil-servizos-emplaza-augas-galicia-actuar-forcadas/20150722220047129867.html

[18] "Cachamuíña se llena de algas tóxicas como las de As Conchas," *La Voz de Galicia*, sep 2011. [Online]. Available: https://www.lavozdegalicia.es/noticia/ourense/san-cibrao-das-vinas/2011/09/17/cachamuina-llena-algas-toxicas-as-conchas/0003_201109O17C10991.htm

[19] "Un "bloom" de cianobacterias tiñe las aguas de as conchas," *La Región*, Jun 2018. [Online]. Available: https://www.laregion.es/articulo/baixa-limia/bloom-cianobacterias-tine-aguas-conchas/20180627223819805235.html

[20] E. Mantzouki and et al., "A European Multi Lake Survey dataset of environmental variables, phytoplankton pigments and cyanotoxins," *Scientific Data*, vol. 5, p. 180226, Oct. 2018. [Online]. Available: https://www.nature.com/articles/sdata2018226

[21] ——, "Temperature Effects Explain Continental Scale Distribution of Cyanobacterial Toxins," *Toxins*, vol. 10, no. 4, 2018.

[22] 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, Mar. 2013. [Online]. Available: https://doi.org/10.1186/1471-2105-14-115

[23] A. Gelzinis, A. Verikas, E. Vaiciukynas, and M. Bacauskiene, "A novel technique to extract accurate cell contours applied for segmentation of phytoplankton images," *Machine Vision and Applications*, vol. 26, no. 2, pp. 305–315, Apr. 2015. [Online]. Available: https://doi.org/10.1007/s00138-014-0643-0

[24] E. C. Orenstein, O. Beijbom, E. E. Peacock, and H. M. Sosik, "WHOI-Plankton- A Large Scale Fine Grained Visual Recognition Benchmark Dataset for Plankton Classification," *arXiv:1510.00745 [cs]*, Oct. 2015. [Online]. Available: http://arxiv.org/abs/1510.00745

[25] H. Zheng, N. Wang, Z. Yu, Z. Gu, and B. Zheng, "Robust and automatic cell detection and segmentation from microscopic images of non-setae phytoplankton species," *IET Image Processing*, vol. 11, no. 11, pp. 1077–1085, 2017.

[26] K. Rodenacker, B. Hense, U. Jütting, and P. Gais, "Automatic analysis of aqueous specimens for phytoplankton structure recognition and population estimation," *Microscopy Research and Technique*, vol. 69, no. 9, pp. 708–720, 2006. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/jemt.20338

[27] E. Álvarez, A. López-Urrutia, and E. Nogueira, "Improvement of plankton biovolume estimates derived from image-based automatic sampling devices: application to

FlowCAM," *Journal of Plankton Research*, vol. 34, no. 6, pp. 454–469, Jun. 2012. [Online]. Available: https://academic.oup.com/plankt/article/34/6/454/1573631

[28] IMDEA Agua, "CIANOALERT: Alerta inteligente contra las floraciones nocivas de cianobacterias." [Online]. Available: https://www.agua.imdea.org/noticias/2017/cianoalert-alerta-inteligente-contra-las-floraciones-nocivas-de-cianobacterias

[29] C. S. Davis, S. M. Gallager, M. Marra, and W. Kenneth Stewart, "Rapid visualization of plankton abundance and taxonomic composition using the Video Plankton Recorder," *Deep Sea Research Part II: Topical Studies in Oceanography*, vol. 43, no. 7, pp. 1947–1970, Jan. 1996. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0967064596000513

[30] C. S. Davis, S. M. Gallager, and A. R. Solow, "Microaggregations of Oceanic Plankton Observed by Towed Video Microscopy," *Science*, vol. 257, no. 5067, pp. 230–232, Jul. 1992. [Online]. Available: https://science.sciencemag.org/content/257/5067/230

[31] Y. Nagashima, Y. Matsumoto, H. Kondo, H. Yamazaki, and S. Gallager, "Development of a realtime plankton image archiver for AUVs," in *2014 IEEE/OES Autonomous Underwater Vehicles (AUV)*, Oct. 2014, pp. 1–6.

[32] H. Sosik and R. J Olson, "Automated taxonomic classification of phytoplankton sampled with imaging-in-flow cytometry," *Limnology and Oceanography: Methods*, vol. 5, 06 2007.

[33] C. A. Graff and J. Ellen, "Correlating Filter Diversity with Convolutional Neural Network Accuracy," in *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, Dec. 2016, pp. 75–80.

[34] A. Kalmbach, Y. Girdhar, H. M. Sosik, and G. Dudek, "Phytoplankton hotspot prediction with an unsupervised spatial community model," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*, May 2017, pp. 4906–4913.

[35] M. G. Camoying and A. T. Yñiguez, "FlowCAM optimization: Attaining good quality images for higher taxonomic classification resolution of natural phytoplankton samples," *Limnology and Oceanography: Methods*, vol. 14, no. 5, pp. 305–314, 2016. [Online]. Available: https://aslopubs.onlinelibrary.wiley.com/doi/abs/10.1002/lom3.10090

[36] E. Álvarez *et al.*, "Routine determination of plankton community composition and size structure: a comparison between FlowCAM and light microscopy," *Journal*

of Plankton Research, vol. 36, no. 1, pp. 170–184, Jan. 2014. [Online]. Available: https://academic.oup.com/plankt/article/36/1/170/1518384

[37] I. Corrêa, P. Drews, M. S. d. Souza, and V. M. Tavano, "Supervised Microalgae Classification in Imbalanced Dataset," in *2016 5th Brazilian Conference on Intelligent Systems (BRACIS)*, Oct. 2016, pp. 49–54.

[38] E. Álvarez, A. López-Urrutia, E. Nogueira, and S. Fraga, "How to effectively sample the plankton size spectrum? A case study using FlowCAM," *Journal of Plankton Research*, vol. 33, no. 7, pp. 1119–1133, Jul. 2011. [Online]. Available: https://academic.oup.com/plankt/article/33/7/1119/1558966

[39] I. Corrêa, P. Drews, S. Botelho, M. S. d. Souza, and V. M. Tavano, "Deep Learning for Microalgae Classification," in *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, Dec. 2017, pp. 20–25.

[40] A. Remsen, "Evolution and field application of a plankton imaging system," Ph.D. dissertation, College of Marine Science, University of South Florida, Jan. 2008.

[41] "Video Plankton Recorder (VPR) - Woods Hole Oceanographic Institution," [Accessed 17-June-2019]. [Online]. Available: https://www.whoi.edu/what-we-do/explore/instruments/instruments-sensors-samplers/video-plankton-recorder-vpr/

[42] "St John's Island National Marine Laboratory newly acquired flowcam 8000 highlights," [Accessed 17-June-2019]. [Online]. Available: http://sjinml.nus.edu.sg/new-equipment-flowcam-8000/

[43] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, June 2009, pp. 1597–1604. [Online]. Available: 10.1109/CVPR.2009.5206596

[44] M. D. Fairchild, "Color Appearance Terminology," in *Color Appearance Models*. John Wiley & Sons, Ltd, 2013, pp. 85–96. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118653128.ch4

[45] M. Kümmerer, T. S. A. Wallis, and M. Bethge, "Information-theoretic model comparison unifies saliency metrics," *Proceedings of the National Academy of Sciences*, vol. 112, no. 52, pp. 16 054–16 059, 2015. [Online]. Available: https://www.pnas.org/content/112/52/16054

[46] A. Verikas, A. Gelzinis, M. Bacauskiene, I. Olenina, S. Olenin, and E. Vaiciukynas, "Phase congruency-based detection of circular objects applied to analysis of phytoplankton

images," *Pattern Recognition*, vol. 45, no. 4, pp. 1659–1670, Apr. 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320311004432

[47] Tong Luo *et al.*, "Recognizing plankton images from the shadow image particle profiling evaluation recorder," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 34, no. 4, pp. 1753–1762, Aug. 2004.

[48] A. McQuatters-Gollop, D. G. Johns *et al.*, "From microscope to management: The critical value of plankton taxonomy to marine policy and biodiversity conservation," *Marine Policy*, vol. 83, pp. 1 – 10, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0308597X16307874

[49] L. Boddy, C. Morris, M. Wilkins, L. Al-Haddad, G. Tarran, R. Jonker, and P. Burkill, "Identification of 72 phytoplankton species by radial basis function neural network analysis of flow cytometric data," *Marine Ecology Progress Series*, vol. 195, pp. 47–59, 2000. [Online]. Available: http://www.int-res.com/abstracts/meps/v195/p47-59/

[50] P. Culverhouse *et al.*, "Automatic classification of field-collected dinoflagellates by artificial neural network," *Marine Ecology Progress Series*, vol. 139, pp. 281–287, Aug. 1996. [Online]. Available: https://www.int-res.com/abstracts/meps/v139/p281-287/

[51] D. A. Lisin, M. A. Mattar, M. B. Blaschko, E. G. Learned-Miller, and M. C. Benfield, "Combining Local and Global Image Features for Object Class Recognition," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Workshops*, Sep. 2005, pp. 47–47.

[52] J. Zhao, H. Guo, and X. Sun, "A Research on the Recognition of Chironomid Larvae Based on SVM," in *2009 Pacific-Asia Conference on Circuits, Communications and Systems*, May 2009, pp. 610–613.

[53] Q. Hu and C. Davis, "Automatic plankton image recognition with co-occurrence matrices and Support Vector Machine," *Marine Ecology Progress Series*, vol. 295, pp. 21–31, 2005. [Online]. Available: http://www.int-res.com/abstracts/meps/v295/p21-31/

[54] K. V. Embleton, C. E. Gibson, and S. I. Heaney, "Automated counting of phytoplankton by pattern recognition: a comparison with a manual counting method," *Journal of Plankton Research*, vol. 25, no. 6, pp. 669–681, Jun. 2003. [Online]. Available: https://academic.oup.com/plankt/article/25/6/669/1553638

[55] H. A. Al-Barazanchi, A. Verma, and S. Wang, "Performance evaluation of hybrid CNN for SIPPER plankton image calssification," in *2015 Third International Conference on Image Information Processing (ICIIP)*, Dec. 2015, pp. 551–556.

[56] E. C. Orenstein and O. Beijbom, "Transfer Learning and Deep Feature Extraction for Planktonic Image Data Sets," in *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Mar. 2017, pp. 1082–1088.

[57] H. Lee, M. Park, and J. Kim, "Plankton classification on imbalanced large scale database via convolutional neural networks with transfer learning," in *2016 IEEE International Conference on Image Processing (ICIP)*, Sep. 2016, pp. 3713–3717.

[58] T. Ojala, M. Pietikainen, and D. Harwood, "Performance evaluation of texture measures with classification based on kullback discrimination of distributions," in *Proceedings of 12th International Conference on Pattern Recognition*, vol. 1, Oct 1994, pp. 582–585 vol.1.

[59] M. R. Turner, "Texture discrimination by gabor functions," *Biological Cybernetics*, vol. 55, no. 2, pp. 71–82, Nov 1986. [Online]. Available: https://doi.org/10.1007/BF00341922

[60] Z. Huang and J. Leng, "Analysis of hu's moment invariants on image scaling and rotation," *Proc. of 2nd International Conference on Computer Engineering and Technology (ICCET)*, vol. 7, pp. V7–476, 05 2010.

[61] H. Zheng, R. Wang, Z. Yu, N. Wang, Z. Gu, and B. Zheng, "Automatic plankton image classification combining multiple view features via multiple kernel learning," *BMC Bioinformatics*, vol. 18, no. 16, p. 570, Dec. 2017. [Online]. Available: https://doi.org/10.1186/s12859-017-1954-8

[62] Xiaoou Tang, "Multiple competitive learning network fusion for object classification," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 28, no. 4, pp. 532–543, Aug. 1998.

[63] A. Verikas, A. Gelzinis, M. Bacauskiene, I. Olenina, and E. Vaiciukynas, "An integrated approach to analysis of phytoplankton images," *Oceanic Engineering, IEEE Journal of*, vol. PP, pp. 1–12, 05 2014.

[64] R. S. Pressman, *Software engineering: a practitioner's approach*, 5th ed., ser. McGraw-Hill series in computer science. Boston, Mass: McGraw Hill, 2000.

[65] Incremental Model in SDLC: Use, Advantage & Disadvantage. [Accessed 06-July-2019]. [Online]. Available: https://www.guru99.com/what-is-incremental-model-in-sdlc-advantages-disadvantages.html

[66] K. Zuiderveld, "VIII.5. - Contrast Limited Adaptive Histogram Equalization," in *Graphics Gems*, P. S. Heckbert, Ed. Academic Press, Jan. 1994, pp. 474–485. [Online]. Available: http://www.sciencedirect.com/science/article/pii/B9780123361561500616

[67] S. Suzuki and K. Abe, "Topological structural analysis of digitized binary images by border following," *Computer Vision, Graphics, and Image Processing*, vol. 30, no. 1, pp. 32–46, Apr. 1985. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0734189X85900167

[68] B. Gärtner and M. Hoffmann, "Computational Geometry Lecture Notes HS 2013," 2013. [Online]. Available: https://www.ti.inf.ethz.ch/ew/Lehre/CG13/lecture/cg-2013.pdf

[69] Y. Zhang, R. Jin, and Z.-H. Zhou, "Understanding bag-of-words model: A statistical framework," *International Journal of Machine Learning and Cybernetics*, vol. 1, pp. 43–52, Dec. 2010.

[70] D. G. Lowe, "Object recognition from local scale-invariant features," *Proceedings of the International Conference on Computer Vision*, vol. 2, pp. 1150–, 1999. [Online]. Available: http://dl.acm.org/citation.cfm?id=850924.851523

[71] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," *Computer Vision-ECCV 2006*, vol. 3951, pp. 404–417, 07 2006.

[72] J. Yousefi, "Image binarization using otsu thresholding algorithm," *University of Guelph*, 05 2015.

[73] R. G. Kuehni, *Color space and its divisions: color order from antiquity to the present.* Hoboken, N.J: Wiley-Interscience, 2003, ch. 2.

[74] SharkD, "RGB-Cube," Mar. 2010, [Accessed 12-July-2019]. [Online]. Available: https://commons.wikimedia.org/wiki/File:RGB_Cube_Show_lowgamma_cutout_b.png

[75] R. C. Niemietz, "English: RGB image split into its three RGB channels," Mar. 2008, [Accessed 12-July-2019]. [Online]. Available: https://commons.wikimedia.org/wiki/File:RGB_channels_separation.png

[76] L. Najman and H. Talbot, "Introduction to Mathematical Morphology," in *Mathematical Morphology.* John Wiley & Sons, Ltd, 2013, pp. 1–33. [Online]. Available: http://onlinelibrary.wiley.com/doi/abs/10.1002/9781118600788.ch1

[77] Gjacquenot, "English: A Delaunay triangulation with circumcircles," Dec. 2013, [Accessed 12-July-2019]. [Online]. Available: https://commons.wikimedia.org/wiki/File:Delaunay_circumcircles_vectorial.svg

[78] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, Sep. 1995. [Online]. Available: http://link.springer.com/10.1007/BF00994018

[79] García-Gonzalo *et al.*, "Hard-Rock Stability Analysis for Span Design in Entry-Type Excavations with Learning Classifiers," *Materials*, vol. 9, p. 531, Jun. 2016.

[80] "Simple Tutorial on SVM and Parameter Tuning in Python and R," Feb. 2017, [Accessed 06-July-2019]. [Online]. Available: https://www.hackerearth.com/blog/developers/simple-tutorial-svm-parameter-tuning-python-r/

[81] N. S. Altman, "An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression," *The American Statistician*, vol. 46, no. 3, pp. 175–185, Aug. 1992. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/00031305.1992.10475879

[82] X. Wu *et al.*, "Top 10 algorithms in data mining," *Knowledge and Information Systems*, vol. 14, no. 1, pp. 1–37, Jan. 2008. [Online]. Available: http://link.springer.com/10.1007/s10115-007-0114-2

[83] M. Kearns, "Thoughts on Hypothesis Boosting." Dec. 1988. [Online]. Available: http://www.cis.upenn.edu/~mkearns/papers/boostnote.pdf

[84] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference and prediction*, 2nd ed. Springer, 2009. [Online]. Available: http://www-stat.stanford.edu/~tibs/ElemStatLearn/

[85] "What is the difference between Bagging and Boosting?" Apr. 2016, [Accessed 17-July-2019]. [Online]. Available: https://quantdare.com/what-is-the-difference-between-bagging-and-boosting/

[86] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, Aug. 1996. [Online]. Available: http://link.springer.com/10.1007/BF00058655

[87] Tin Kam Ho, "Random decision forests," in *Proceedings of 3rd International Conference on Document Analysis and Recognition*, vol. 1, Aug 1995, pp. 278–282. [Online]. Available: 10.1109/ICDAR.1995.598994

[88] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, "Numerical Recepies: The Art of Scientific Computing," 2007.

[89] F. D. College and F. Dellaert, "The Expectation Maximization Algorithm," Tech. Rep., 2002.

[90] scikit-learn documentation, "Gaussian mixture models," [Accessed 12-July-2019]. [Online]. Available: https://scikit-learn.org/stable/modules/mixture.html