Semi-Automatic Segmentation, Detection and Classification of Gram Stained Bacteria in Blood Samples

Sofia Lejon, Emelie Andersson

Master's thesis
2016:E23

Lund University

Faculty of Engineering
Centre for Mathematical Sciences
Mathematics

Centrum Scientiarum Mathematicarum

Abstract

Manual microscopy is a time-consuming and inefficient procedure in microbiology laboratories today. Common analyses in these laboratories are detection and classification of Gram stained bacteria [1]. Bacteria that have been Gram stained are either Gram negative or Gram positive. Gram negative bacteria are pink/red and Gram positive bacteria are purple. Moreover, the bacteria can have different morphology. The most common are cocci and bacilli, cocci are round and bacilli are rod-shaped bacteria. Lastly, they are classified based on how they grow which, for instance, can be in chains or clusters.

This thesis investigates whether it is possible to make an automatic, digital system that can replace manual microscopy for Gram stained bacteria. Images of bacteria were acquired with a digital microscope, provided by the company where the thesis was written, CellaVision AB.

A method that segmented bacteria from background in the images was developed. Moreover, several methods have been implemented aiming to detect and classify bacteria based on their color, shape and arrangement.

A final system was created that combined the most successful methods that enabled detection and classification of Gram stained bacteria. It could be concluded that an automatic, digital system for detection and classification of Gram stained bacteria is possible to implement. The system developed in this thesis was however semi-automatic since some user input was needed.

Preface

This report is our Master's thesis for the degrees in Master of Science in Biomedical Engineering. The thesis was written at the company CellaVision AB during 20 weeks in Spring 2016. The thesis was performed for the Centre for Mathematical Sciences at Lund Institute of Technology (LTH).

First of all we would like to thank our supervisors Niels Christian Overgaard from Centre at Mathematical Sciences at LTH and Kenth Stråhlén from Cellavision AB. We would especially like to thank Niels Chrsitian Overgaard for your guidance regarding mathematical and image analysis concepts and Kenth Stråhlén for your support and help in everyday matters.

Finally, we want to thank all the employees at CellaVision AB that have helped us and made us feel as a part of the company.

Contents

Chapter 1

Introduction

In this chapter the aim of the thesis, general microbiology and bacteriology will be introduced and explained.

1.1 Aim of the Thesis

The aim of this thesis is to investigate if it is possible to automatically (or semi-automatically) detect Gram stained bacteria in images and classify them. The bacteria should be classified based on their color, shape and arrangement. The color is classified as Gram negative (pink/red) or Gram positive (purple). If the shape is round the bacteria are classified as cocci and if they are rod-shaped as bacilli. Lastly, the bacteria should be classified depending on their arrangement which for instance can be that they grow in a chain or in a cluster.

1.2 CellaVision AB

The master thesis was written at the company CellaVision AB. CellaVision develops systems for automatic pre-classification and visual presentation of blood cells for laboratories all over the world. Both the complex hardware, which includes high magnification optics, camera and precision robotics, and the software, which includes image analysis and a user-interface, are developed at the company. CellaVision's system has many different applications and the portfolio keeps growing. This master's thesis is an initial attempt for CellaVision to enter the world of microbiology.

1.3 Microbiology

The work flow, analysis and procedures can vary extensively between different microbiology laboratories. Traditionally, and still in some smaller laboratories, no part of the work flow is digitalized. However, digital pathology, which is the concept of handling, analyzing and storing pathology slides digitally, is an emerging market. As part of the information collection for the background study a visit to the microbiology laboratory in Lund was carried out to investigate whether there are any analyses in microbiology that are suitable for digitalization and automatization. The purpose of automatic digital microscopy is to generate a faster and more standardized analysis. In addition, other positive aspects of digital microscopy are that the images can be stored and viewed from another location and that less labor is required to perform the analysis. As a result, digital microscopy can both improve the performance of the analysis and reduce the costs.

In order for an analysis to be suitable for automatization certain qualities need to be fulfilled. According to Johan Rydberg, chief physician at the Microbiology department at Lund University hospital, several samples need to be analyzed every day if an analyzing product is to be profitable for the lab to invest in. About 15 000 bacteria samples are looked at in a microscope each year at the Microbiology department at Lund University hospital which is enough for an investment to be interesting [1]. Another crucial aspect is that the system, consisting of a digital microscope and software, is able to perform the same analysis as a human. Considering these aspects and the information gathered from the laboratory in Lund, classification of bacteria was decided to be the most promising business out of the microbiology portfolio.

When a bacteria infection is suspected, samples from the patient are collected. These samples can vary and can for instance be spinal fluid, blood, pleural fluid (from the lungs) or biopsies [1]. Blood samples were recognized as the most suitable analysis to be automatized in microbiology. The first reason for choosing blood is that there are more blood samples analyzed compared to samples from other body fluids. The second reason is the importance to get answers quickly when there are bacteria in the blood, which is not always achieved now as most of the laboratories do not have staff that can analyze the samples during the night [1].

There are generally few bacteria in a sample and the chances of detecting them in a microscopy are very small. To ensure the amount of bacteria is

enough for analysis, the bacteria are cultivated. This cultivation takes several hours to execute and is also done during the night. When the bacteria have been cultivated long enough an alarm is triggered and the laboratory personnel stop the cultivation. Since there are no laboratory personnel that can make the analysis at night it results in that some samples are not analyzed until the morning. An automatic system could provide a microscopic analysis both day and night.

After the bacteria sample has been cultivated it has to be stained to make the bacteria visible in a light microscopy. The most common staining for bacteria is Gram staining which will be described in greater detail below in section 1.4.1. Gram stain is used in practice to get a fast indication of what type of bacteria the sample contains. The traditional way to get a more detailed indication of the actual bacterial species is to cultivate in different mediums, such as nutrient broths [1]. However, new technology is disrupting the procedure where the species are determined. The most popular technique is Maldi-tof which is a mass spectrometry method used to find bacterial species. It is for example used at the laboratory in Lund. A more precise classification is important since it can provide a more effective antibiotics treatment. The reason that Gram staining is still done as a first rough classification is due to Maldi-tof's longer execution time. Consequently, Gram staining is still an important analysis for bacterial identification.

Many diseases can be diagnosed by symptoms and visual inspection of the patient. However, the symptoms of a bacterial infection are ambiguous and can in most cases not be distinguished from a viral infection [2]. The use of microscopy to detect bacterial infections are therefore of great importance.

1.4 Bacteria

1.4.1 Gram Staining

When an infection is suspected body fluids and biopsies are collected to make an analysis to investigate whether or not bacteria are present. In order to see the bacteria in an optical microscope they need to be attached to a glass slide and then stained. Figure 1.1 shows such a slide. The glass slide with the blood sample is called a blood smear. If the sample is not stained it will be hard to detect the bacteria since they are basically colorless without the staining [3]. The most common coloring method for bacteria is Gram staining [4].

Figure 1.1: An example of a bacterial smear on a glass slide that has been Gram stained.

Gram staining differentiates Gram positive bacteria from Gram negative bacteria, which is a way of dividing different bacterial species into two large groups. The Gram positive bacteria are purple and the Gram negative are pink/red after the staining procedure. The sample, containing bacteria, are put on a glass slide. The slide is then heated with a burner to get rid of excess fluid and to fixate the bacteria. In the next step the stain Crystal Violet is added which gives the bacteria the purple color. Iodine is then dropped on the sample to create larger molecules together with the Crystal Violet which further fixates the sample [4]. After this, alcohol or acetone is poured onto the sample to wash away the color from the Gram negative bacteria. The Gram negative bacteria are then colored with a counterstain, most commonly Safranin [3].

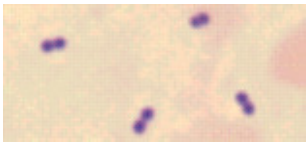
The physical properties that makes it possible for the differentiation between the two types of bacteria are their cell walls which have different thickness in the peptidoglycan layer. The alcohol dehydrates the peptidoglycan cell wall and for the Gram positive bacteria the thick peptidoglycan layer tightens. Since the Crystal Violet and the Iodine solution creates such large molecules they cannot escape through the membrane. The thin cell wall of the Gram negative bacteria is more affected by the alcohol which creates cracks for the color to get washed off. The counter-stain is then applied to make the Gram negative bacteria pink and thus visible in a light microscope.

Gram staining is a relatively quick method to color the bacteria in a sam-

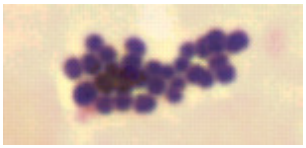
ple. As mentioned earlier, that is an important quality since early detection is crucial from a clinical point of view and the method is therefore widely used.

1.4.2 Bacterial Morphology

Apart from the staining there are other properties to consider for diagnosis of bacteria, namely its morphology and arrangement. The most common morphologies of bacteria are shapes that are round, rod-like and elliptical. The round bacteria are called Cocci, the rod-like are called Bacilli and the elliptical are called Coccobacilli. Apart from these shapes there are corkscrew forms, helical forms and a variety of other shapes which are not as common as the three previously mentioned. The shapes mentioned are also a part of the diagnostic analysis. The different types of arrangement of bacteria are singles, pairs, chains or clusters [5]. Bacteria growing in clusters are called staphylo-like bacteria, some examples are Staphylococcus, Micrococcus and Stomatococcus. Bacteria growing in chains are called strepto-like bacteria, some examples are Streptococcus, Enterococcus and Gemella. The different types of bacteria listed above are all Gram positive cocci, but bacilli can grow in those formations as well. Additionally, the rod shaped bacteria can also grow in palisading and filamentous arrangements. Palisading is when the bacteria creates a chain connecting with the wider end instead of the short so that they lie parallel to each other. One example of such bacteria is Listeria. A filamentous arrangement is when the rods branch out in chains of different length. One example of this is Actinomyces [6]. Figure 1.2 shows examples of different arrangement, colors and shapes of bacteria from samples used in this thesis.

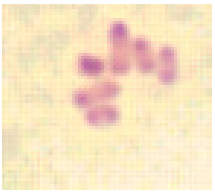


(a) Gram positive cocci in pairs



(b) Gram positive cocci in cluster

(c) Gram positive cocci in chains



(d) Palisiding Gram negative bacilli

Figure 1.2: Examples of different stains, morphology and arrangement of bacteria.

Chapter 2

Data

The data that has been used to test the algorithms were given from the microbiology laboratory in Lund, the company Copan and another microbiology laboratory in Calgary, Canada. All the samples consist of smears on glass microscope slides stained with Gram stain, such as the one shown in Figure 1.1. The images were taken with the CellaVision®DM1200 system.

2.1 Image Acquisition

Bacteria are in the size range of 0.5 - 2 micrometers [6]. To be able to see the bacteria clearly a magnification of 100 x is needed. In reality this is a magnification of 1000 x, since the 'eyepiece' has magnification 10 x. However, it is more common to only mention the magnification of the objective. An objective of 100 x is close to the largest magnification a light microscope can have before diffraction occurs and limits the resolution.

The CellaVision ®DM1200 was used to acquire the images that were analyzed in this thesis. The camera is a Basler FireWire camera and takes color images in 658×492 pixels resolution. The objective is an oil immersion Olympus PLCN 100x objective.

2.2 Samples from Lund

The samples were collected in February 2016. The sample set consisted of ten different glass slides with blood samples containing bacteria. The blood contained different bacteria and some samples had a combination of two

types of bacteria. The bacteria present in the samples are Staphylococcus epidemidis, Eschericha coli, Enterococcus faecium, Streptococcus pneumoniae, Staphylococcus aureus, Streptococcus anginosus, Klebsiella pneumoniae, Escherichia coli and Streptococcus pyogenes. In these samples all of the cocci were Gram positive and all of the bacilli were Gram negative.

2.3 Samples from Copan

The samples were sent from the Italian company Copan in 2012. There were 35 glass slides collected from various body substances such as blood, pleural fluid and kidney fluid. The non-blood samples were discarded since blood samples were the only samples of interest in this case. The type of bacteria were not given.

2.4 Samples from Calgary

The samples from the laboratory in Calgary were also sent in 2012. The set consisted of 10 different samples from various body substances including for example blood and kidney fluid. The samples contained following bacteria: Streptococcus pneumoniae, Escherichia coli, Staphylococcus aureus, Enterococcus faecalis, Haemophilus influenzae, Haemophilus parainfluenzae and Pseudomonas aeruginosa.

2.5 Bacterial species present in the samples

There were many different bacterial species available in the samples. The different species exhibit different characteristics, have different habitats and different effect on the body. Some of the bacteria are a normal part of the human flora. Such bacteria can however cause infections when they enter a part of the body they normally don't inhabit.

2.5.1 Gram positive bacteria

Staphylococcus epidermidis is a coccus bacterium. Most often it can be found on the skin where it is a part of the body's normal flora, yet it can cause infection in combination with for example invasive medical devices [7].

Enterococcus faecium is a bacterium which has a morphology that can vary between spherical and ovoid. This bacterium can be found in the intestine where it is harmless. However it can cause food poisoning from for example unpasteurized milk [8].

Streptococcus pneumoniae and Streptococcus anginosus are two examples of Strepcoccus naturally found in the body. Streptococcus pneumoniae is, as can be guessed by its name, a type of cocci bacteria. It can be found in the nasopharynx, which is an area between the nasal cavity and the esophagus. S. pneumoniae can cause sinus infection, infection in the respiratory tract and infection in the middle ear. It can, in worst cases, lead to deadly infections and is a fairly common death cause for small children in developing countries [9]. Streptococcus anginosus is a type of cocci bacteria that cause infections that leads to abscesses in liver, lung and brain [10].

Other bacteria have habitats outside the body and enter by for example food and wounds. One example is Staphylococcus aureus which is a coccus bacterium. It can be found in soil, water and air and can cause food poisoning when exposed to orally [8]. Another example is Streptococcus pyogenes which also is a coccus bacterium. It enters the body orally, and for those who get an infection the symptoms are often sore throat that can escalate to nausea, fever and vomiting [8].

2.5.2 Gram negative bacteria

A bacteria that can be harmless when it is in the right place in the body is the common Escherichia coli. It is a rod-shaped bacteria most often found in the intestines but can cause diarrhea when exposed to orally. It can also cause urinary tract infection [11]. Klebsiella pneumoniae is a rod-shaped bacterium. It can be found in many places of the body such as the mouth, skin and intestine. The bacterium can cause gastroenteritis for some people, however most people do not become sick of this bacteria [8]. Haemophilus influenzae is another type of bacteria that can live in the host without causing infections. Haemophilus is a coccobacilli. If the immune system is decreased or the host has another infection such as a viral one, it can attack the host. It can cause infections in the upper respiratory tract, pneumonia and bronchitis. It does not cause, as the name might suggest, influenza [12]. A close relative to Haemophilus influenzae is Haemophilus parainfluenzae which can be the source of endocarditis, meningitis and bacteremia (bacteria in the blood) [13].

Chapter 3

Methods and Theory

This chapter describes the methods used to achieve the detection and classification of the bacteria in the images as well as the theory the methods are based on. The methods implemented are aiming to achieve all the different steps described in the flowchart in Figure 3.1.

In the last part of the chapter there is a description of how the final system was put together and which methods that were selected for the final detection and classification system.

All of the methods are implemented in the mathematical computing and programming software MATLAB. For some methods an already existing function from a MATLAB toolbox has been used. In those cases the name of the MATLAB-function has been written.

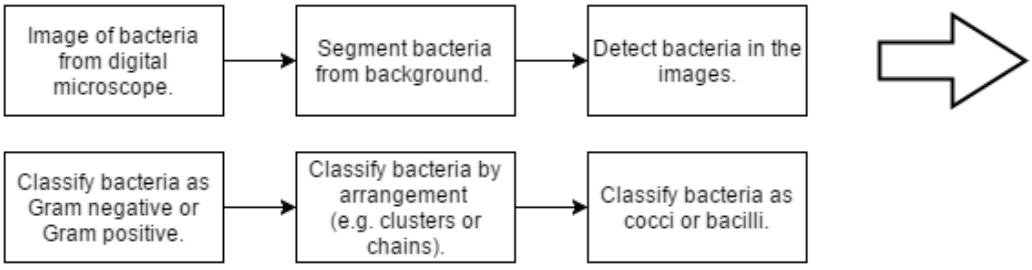


Figure 3.1: Flowchart of the different steps the microscopic image with bacteria will go through.

3.1 The Images Used

The size of the images acquired from the CellaVision camera are 480x640 pixels large and saved in the format of bmp (bitmap image file).

The images used were RGB color images. RGB stands for red, green and blue and refers to one of the possible ways to represent a color image. In digital form an RGB image is a three layer matrix of dimension $m \times n \times 3$ where m and n is the size of the image. Each pixel has a value from each of the three colors which are combined resulting in the final color that is shown in the image. For these images, each value ranged from 0 to 255. Each of the layers red, green and blue are referred to as color channels.

3.2 Stain Normalization

A common problem in analysis of stained samples is inconsistencies in the preparation and handling of the samples. Although the staining method is the same, for instance Gram staining, the visual properties can vary due to these inconsistencies which most often are expressed in differences in colors and intensity. According to Macenko et al., differences in color can for instance depend on how much stain that is added, which type of stain is used or if the sample has been exposed to light which can cause fading of color [14]. Macenko et al. have developed a method to compensate for these factors, called stain normalization. To cope with staining inconsistencies in the samples used in this thesis, the method described by Macenko et al. have been altered to suit the bacteria images.

3.2.1 The model and the stain vectors

Since the stained objects are the only objects of interest in the image the two staining colors in the image make up all the information needed to describe the interesting parts of the image and a model of the image can therefore be explained as:

$$I^* = \boldsymbol{s_1}\alpha_1 + \boldsymbol{s_2}\alpha_2, \tag{3.1}$$

where I^* is the modeled image, $\boldsymbol{s_1}$ and $\boldsymbol{s_2}$ are two RGB-vectors describing the two stains, crystal violet and saffarin and α_1 and α_2 are matrices of the same size as the image describing the saturation of each stain of every pixel in the image. In order to normalize the stains and make the staining colors

similar in every image, which was the goal with this method, a series of steps needed to be carried out. First of all the different stains (Gram positive and Gram negative) needed to be found in the image, the two stains in each image are referred to as stain vectors. Once the stain vectors are found, the saturation of each stain in the image (α_1 and α_2) can be determined.

Two different approaches were investigated to find the two stain vectors; one where the stains were manually chosen and a second where they were found automatically. Moreover, a version for handling only a single stain in the image was implemented for the manual version.

3.2.2 Conversion to the Optical Density space

For both approaches, the image was first transformed to the so called Optical Density space (OD-space) to achieve a greater separation of the colors i.e. a bigger difference between Gram positive and Gram negative color. The image that is transformed is a normalized image where each pixel has a value between 0 and 1 in each color channel. A conversion to the OD-space is done in the following way: $\text{OD} = -\log(I_k)$, where I_k is the respective color channel:

I_1 = Red color channel
I_2 = Green color channel
I_3 = Blue color channel

The pixels below 0.15 in the OD space were assumed to be background. All the background pixels were assigned a value of 0. The remaining parts of the image contain Gram stained bacteria and/or other substances, such as white blood cells, which also can get affected by the staining.

3.2.3 Automatic selection of stain vectors

For the automatic approach, the data set was reduced to two dimensions so that the pixels were located in a plane, rather than a three-dimensional cloud. The plane was created by an orthonormal base which was found by calculating the two largest eigenvectors of the data cloud using Singular Value Decomposition (SVD). The purpose of SVD is to find the factorization of a real or a complex matrix. By doing SVD a data set X can be decomposed

as:

$$X = U\Sigma V^{\top}. \tag{3.2}$$

If X is a real matrix, which it is in this case, U and V are orthogonal matrices that satisfy [15]

$$U^{\top}XV = \Sigma = diag(\sigma_1, ..., \sigma_n). \tag{3.3}$$

The decomposition can also be written in the following way:

$$X = \sigma_1 \mathbf{U}_1 \mathbf{V}_1^{\top} + \sigma_2 \mathbf{U}_2 \mathbf{V}_2^{\top} + ... + \sigma_n \mathbf{U}_n \mathbf{V}_n^{\top}. \tag{3.4}$$

The singular values σ_i are ordered as follows $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_n \geq 0$ [16]. The larger the singular value is the more variance it corresponds to, and hence more information. Consequently, the first singular values are the most important to use to describe the data and the dimensionality of the data set can be reduced by keeping the strongest singular values. The pixels, represented by vector $\boldsymbol{u}$, were then projected down onto the spanning set created by the two most dominant principal components, $\boldsymbol{V_1}$ and $\boldsymbol{V_2}$ in the following way:

$$\boldsymbol{P_u} = (\boldsymbol{V_1} \cdot \boldsymbol{u})\boldsymbol{V_1} + (\boldsymbol{V_2} \cdot \boldsymbol{u})\boldsymbol{V_2}. \tag{3.5}$$

Where $\boldsymbol{P_u}$ is the RGB vector of the pixel projected onto the plane. An illustration can be seen in Figure 3.2. The reason for the projection onto

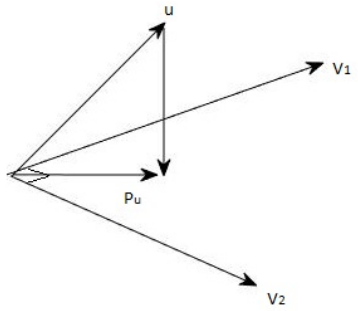


Figure 3.2: An illustration of the projection of pixel vector $\boldsymbol{u}$ onto the plane spanned by $\boldsymbol{V_1}$ and $\boldsymbol{V_2}$.

the plane was to enable a calculation of the angle between every projected pixel ($\boldsymbol{P_u}$) of the new data on the plane and the first eigenvector ($\boldsymbol{V_1}$), which is also a part of the plane. The angles can be interpreted as a key to the

different colors since every pixel color corresponds to a certain angle. The angle between a projected pixel and the first eigenvector was calculated in the following manner:

$$cos(\theta_i) = \frac{\boldsymbol{P_u} \cdot \boldsymbol{V_1}}{\|\boldsymbol{P_u}\|\|\boldsymbol{V_1}\|}. \tag{3.6}$$

Where θ_i is the angle, $\boldsymbol{P_u}$ is the projected pixel vector corresponding to the angle and $\boldsymbol{V_1}$ is the first eigenvector. After the angles of every pixel were calculated they were studied in a histogram. The goal was to find two angles that represented the different stains. The angles in between are assumed to be a linear combination of the two stains.

To make the procedure robust to outliers the 2nd and 98th percentile of the angles were used to find the two stain vectors. Outliers are values in a data set that differ from the rest of the data. They can affect modeling and other methods negatively and are therefore often chosen to be ignored. When the two angles were chosen, their respective projected pixel was chosen as stain vector $(\boldsymbol{s_1}, \boldsymbol{s_2})$, and was thus an RGB-vector describing the color of respective stain in the current image.

The stain vectors are used to reduce the dimensionality, however they are not orthonormal to each other. Using the stain vectors is a minimal way of describing the data since it was assumed that every pixel between the stain vectors can be described as a linear combination of the two stain vectors.

An illustration of the pixels projected onto the plane spanned by $\boldsymbol{V_1}$ and $\boldsymbol{V_2}$ can be seen in Figure 3.3. In the figure, the two stain vectors, $\boldsymbol{s_1}$ and $\boldsymbol{s_2}$ are also illustrated.

3.2.4 Manual selection of stain vectors

In the manual method the stain vectors $(\boldsymbol{s_1}, \boldsymbol{s_2})$ were chosen from the image by a user. The user decided which color that represented stain 1 (i.e. Gram positive) and stain 2 (i.e. Gram negative) in every image. The decision was made by first clicking in the image where Gram positive stain was displayed, thereafter the user clicked on Gram negative stain. The RGB-vectors where the user has clicked are stored as stain vectors $(\boldsymbol{s_1}, \boldsymbol{s_2})$.

3.2.5 Color deconvolution

Once the stain vectors were chosen by either using the automatic or the manual approach, a color deconvolution was performed to find the saturation

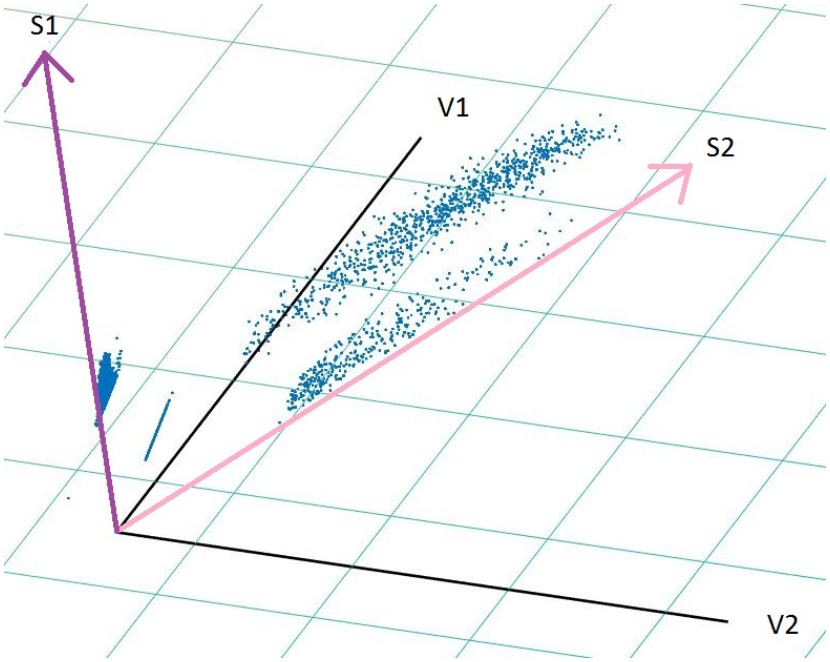


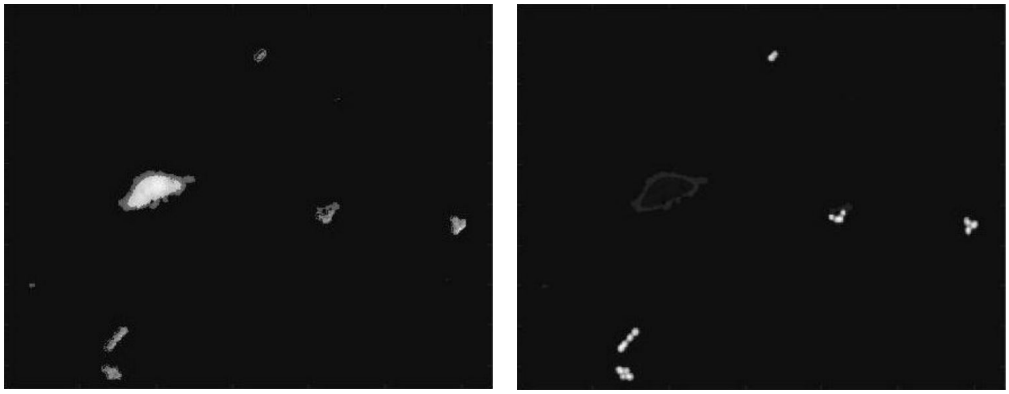
Figure 3.3: The projected pixels in the plane spanned by the first and second eigenvectors ($\boldsymbol{V_1}$ and $\boldsymbol{V_2}$). The two stain vectors $\boldsymbol{s_1}$ and $\boldsymbol{s_2}$ are also illustrated. The pixels in between can be represented as a linear combination of the two stain vectors.

(α_1 and α_2) of each of the stains in the model shown in equation 3.1[17]. Following formula was used to retrieve the two saturation matrices:

$$A = P_{im}S^{-1}. \tag{3.7}$$

A is a matrix containing the saturations of the two stains, α_1 and α_2, and will take real values between 0 and 1 for every pixel in the image. S is a matrix with the two stain vectors, $\boldsymbol{s_1}$ and $\boldsymbol{s_2}$. Lastly, P_{im} is a matrix with the projected pixels in the plane. This resulted in two intensity images, one image that corresponded to the saturation of Gram positive stain (α_1) and another that corresponded to Gram negative stain (α_2), an example of intensity images can be seen in Figure 3.4. A bright pixel corresponds to a high saturation of the stain.

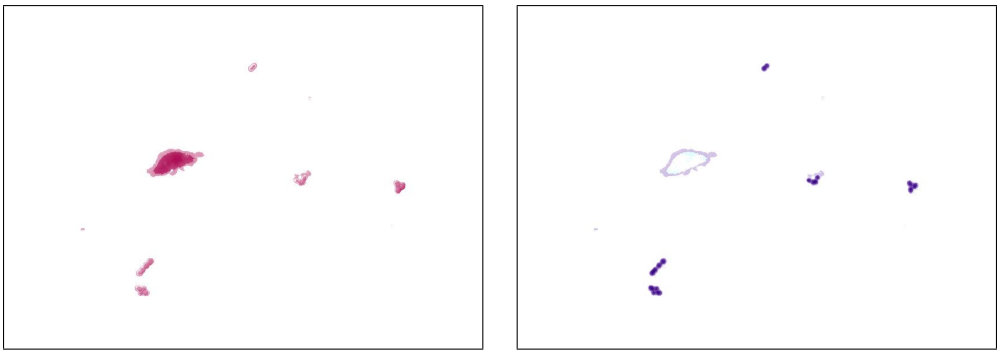
The saturation images can then be multiplied with a standard Gram positive ($\boldsymbol{s_1^*}$) and Gram negative ($\boldsymbol{s_2^*}$) colors that have been chosen manually from an image with representable stain colors, an example of the result is



(a) Saturation of Gram negative stain. (b) Saturation of Gram positive stain.

Figure 3.4: Saturations of stains.

found in Figure 3.5a and Figure 3.5b.



(a) Colored with Gram negative stain. (b) Colored with Gram positive stain.

Figure 3.5: The intensity images colored with their respective stain.

As mentioned before the image can be modelled according to eqution 3.1. To retrieve the resulting stain normalized image a slightly modified version was used:

$$RGB_{OD} = \alpha_1 \boldsymbol{s_1^*} + \alpha_2 \boldsymbol{s_2^*}. \tag{3.8}$$

RGB_{OD} is the color image matrix in the OD-space recreated with the intensity images and the standard stains used to color the image. Where α_1 is the saturation matrix of stain 1, $\boldsymbol{s_1^*}$ is the standard color vector of stain 1, α_2 is the saturation matrix of stain 2 and $\boldsymbol{s_2^*}$ is the standard color of stain 2.

However, before the images were stained the intensity needed to be corrected since it can vary among different images. Two maxima were chosen, one for each stain. These maxima, which are called pseudo-maxima, are the intensity values all images are normalized to. The values of the pseudo-maxima were collected from the image where the standard stain vectors ($\boldsymbol{s_1^*}$ and $\boldsymbol{s_2^*}$) used to stain the image were found. The reason for choosing that image was because it had a satisfactory intensity level, which was desirable to achieve with the other images as well. The intensity images were then normalized and multiplied with the pseudo-maxima so that every image had the same intensity range. Finally, the normalized stains could be combined to form a resulting stain normalized image, using Equation 3.8. The image was then transformed back to the original color space from the OD-space using following relationship:

$$I_k = 10^{(-RGB_{OD})}, \tag{3.9}$$

where I_k is the reconstructed image after the stain normalization in the original color space and RGB_{OD} is the color image matrix in the OD-space.

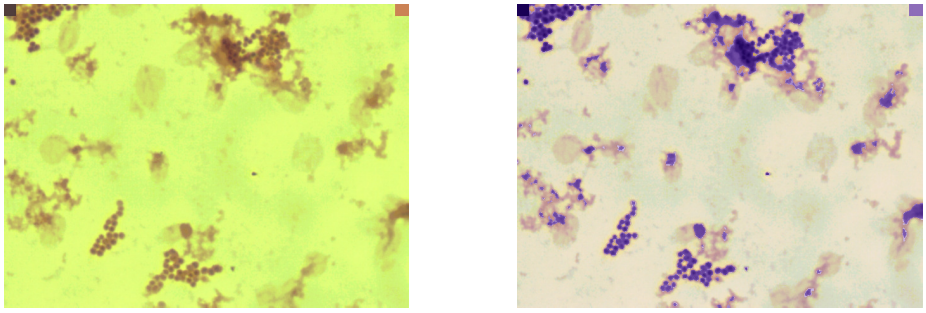
3.2.6 Manual selection handling a single stain

There were several images where only one stain was present, such as only Gram positive bacteria. To cope with that the manual method was extended to also handle images with only one stain.

The user is asked to first click on Gram positive stain in the image and thereafter the Gram negative stain. If there is no Gram positive stain in the image the user instead clicks on a Gram negative square in the upper left corner, see Figure 3.6a. When only Gram negative is present the user instead clicks on the Gram positive colored square in the right upper corner.

The algorithm then finds the saturation of the stain and colors the image with the standard color as described in section 3.2.5. The rest of the image is treated as background.

An example of the result can be seen in 3.6b with the original image in Figure 3.6a, the background has also been altered to mimic the background in a light microscope.



(a) The original image. (b) A stain normalized image.

Figure 3.6: A comparison of an image before and after stain normalization.

3.3 Segmentation

Segmentation is the division of an image into meaningful structures and is often an essential step to enable further analysis of an image. Meaningful structures are the regions of interest in each particular image. There are a few different ways to segment interesting objects in an image. A widely used segmentation method is image thresholding which will be explained in further depth below. Other examples are edge based segmentation, region based segmentation, clustering techniques and matching [18].

3.3.1 Image thresholding

Image thresholding is a straightforward way to segment regions of interest in an image. In thresholding a gray scale image is converted to a binary image using a specific threshold. The threshold is chosen so that an optimal segmentation of the interesting regions from the background is achieved. Let $I_{i,j}$ be a pixel in a gray scale image, let $B_{i,j}$ be a pixel in a binary image and t be the threshold. Then the tresholding is done in the following way:

$$B_{i,j} = \begin{cases} 1 & \text{if } I_{i,j} \geq t \\ 0 & \text{if } I_{i,j} < t \end{cases} \tag{3.10}$$

3.3.2 Otsu's method for image thresholding

Otsu's method for image thresholding was presented in 1979 and has since then been widely used in computer vision and image processing [19]. The method aims to find an optimal threshold value to extract interesting objects in the image from the background.

It is of interest to automatically find the threshold described in equation 3.10. To do so the histogram of the image is often used. In an ideal case, there is a distinct difference in the histogram between foreground and background expressed as a valley. In practice, however, it is often hard to find the valley separating interesting objects from the background. To overcome this problem, Otsu evaluated the goodness of the chosen threshold by investigating the variance within the separated classes. The optimal threshold minimizes the intra-class variance which is defined as a sum of weighted variances. When separating pixels in an image into two classes the sum of the weighted variances are defined as:

$$\sigma_W^2(t) = \omega_1(t)\sigma_1^2(t) + \omega_2(t)\sigma_2^2(t). \tag{3.11}$$

Where $\sigma_1^2(t)$ is the variance of class 1 (foreground) and $\sigma_2^2(t)$ is the variance of class 2 (background). ω_1 and ω_2 are the probabilities of respective class. The probabilities have been calculated from the histogram. The histogram has been divided into two histograms by the threshold that is to be evaluated. If there are L pixels, the first histogram contains pixel values from 1 to t and the second pixels from t+1 to L, where t is the threshold. The probabilties are calculated from the histograms in the following way:

$$\omega_1(t) = Pr(C_1) = \sum_{i=1}^{t} p_i = \omega(k). \tag{3.12}$$

$$\omega_2(t) = Pr(C_2) = \sum_{i=t+1}^{L} p_i = 1 - \omega(k). \tag{3.13}$$

Where C_1 and C_2 are two different classes and p_i is the probability distribution. The mean intensity levels of the two classes are defined as:

$$\mu_1(t) = \sum_{i=1}^{t} \frac{ip_i}{\omega_1}. \tag{3.14}$$

$$\mu_2(t) = \sum_{i=t+1}^{L} \frac{ip_i}{\omega_2}. \tag{3.15}$$

The mean intensity of the total image is defined as:

$$\mu_T = \sum_{i=1}^{L} ip_i. \tag{3.16}$$

Otsu showed that maximizing the inter-class variance (the variance between classes) is the same as minimizing the intra-class variance. The inter-class variance σ_B^2 is defined as the difference between the total variance of the image σ_T^2 and the variance within classes σ_W^2:

$$\sigma_B^2 = \sigma_T^2 - \sigma_W^2 = \omega_1(\mu_1 - \mu_T)^2 + \omega_2(\mu_2 - \mu_T)^2. \tag{3.17}$$

The optimal threshold t^* is then the value that maximizes σ_B^2:

$$\sigma_B^2(t^*) = \max_{1 \leq t \leq L} \sigma_B^2(t). \tag{3.18}$$

3.3.3 Segmentation Using Dempster Shafer

Segmenting objects of interest can be a challenging problem when working with microscopic images due to the limited resolution of light microscopes. It is especially hard when the objects in the image are only a few micrometers wide, like bacteria. Additionally, it is even harder to segment bacteria that have faint color, such as Gram negative bacteria. The problem lies in choosing an adequate thresholding value since the boundary between the bacteria and the background in many cases are unclear. The unclear boundary of a Gram negative bacterium is visualized in Figure 3.7.

In MATLAB there is an already existing implementation of Otsu's thresholding method as a function named `graythresh`. The function was tested but proven to be insufficient for these kind of images. To solve the thresholding problem, the properties that the images were stained were used and a more advanced color thresholding method was implemented.

An image is traditionally represented in the RGB color space. However, there are several other color spaces to describe an image with. Some examples of color spaces are HSI, LAB, and YCbCr. As can be seen in Figure 3.8 different content of the image is emphasized depending on which color space the image is represented in.

Since the color spaces represent the colors in different ways they also highlight different information in the image. These differences were used to separate the bacteria from the background. The intention was to find

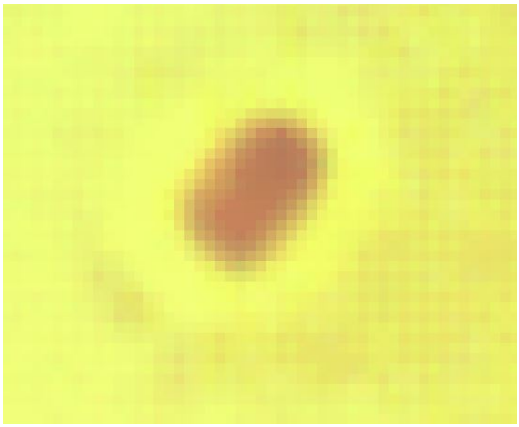
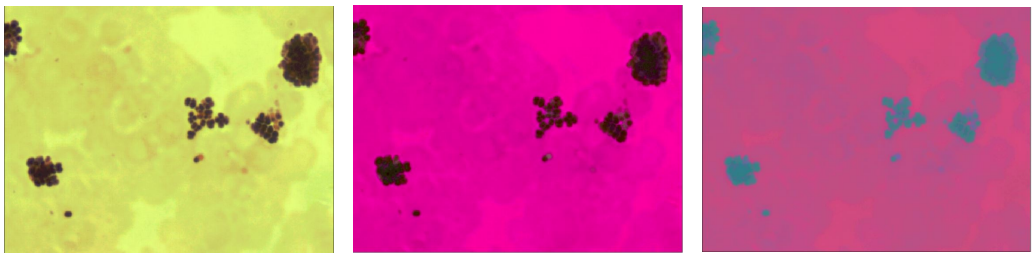


Figure 3.7: A Gram negative bacterium captured with a light microscope at 100x resolution.



(a) RGB (b) LAB (c) YCbCr

Figure 3.8: An image represented in (from left to right): RGB, LAB and YCbCr.

the most sensitive color channel in each respective color space and make an initial segmentation (using Otsu) of foreground (in this case bacteria) and background. Different color channels will have different judgement on whether a pixel is bacteria or not. The idea is to then merge the different judgements and thus use information from several sources to decide if a pixel is foreground or background. To accomplish this a statistical method called Dempster Shafer (DS) was used. The work presented here is based on Harrabi et al.'s method for using DS segmentation on microscopic medical images [20]. DS is able to combine different information and provide a final segmentation of the bacteria. The overall term of techniques using data from different sources and combining them are called data fusion techniques and DS is an example of that. Data fusion methods are mostly useful when

information is incomplete, imprecise or uncertain, like in this case where it is hard to determine whether a pixel is foreground or not. The drawback of these techniques is that they can be computationally heavy [20].

3.3.3.1 Conversion between color spaces

RGB is one of many color spaces that can be used to represent a color image. The image can be converted from the RGB color space to another color space and the conversion can be both linear and non-linear. A non-linear transformation is assumed to contribute with new information about the color of the image. Below is an example of a non-linear conversion where an RGB-image is converted to the Hue Saturation and Intensity (HSI) color space.

The RGB-values are normalized to have values between 0 and 1 and this conversion results in normalized values in the HSI-space as well.

$$h = \cos^{-1} \frac{0.5((r-g)+(r-b))}{((r-g)^2+(r-g)(r-b))^{\frac{1}{2}}}, \quad h \in [0,\pi], \text{ for } b \leq g, \tag{3.19}$$

$$h = 2\pi - \cos^{-1} \frac{0.5((r-g)+(r-b))}{((r-g)^2+(r-g)(r-b))^{\frac{1}{2}}}, \quad h \in [\pi,2\pi], \text{ for } b > g, \tag{3.20}$$

$$s = 1 - 3\min(r,g,b), \quad s \in [0,1], \tag{3.21}$$

$$i = \frac{r+g+b}{3}, \quad i \in [0,1]. \tag{3.22}$$

Where r is the normalized value of the pixel from the red channel, g is the normalized pixel value from the green channel and b is the normalized pixel value from the blue channel. The resulting components h, s and i are the normalized values used in the HSI color space [21].

3.3.3.2 The color spaces used

The color spaces used in this application were: RGB, HSI, XYZ, YCbCr, LAB, LUV and YIQ. The majority of the color spaces used were not linear transformations of RGB (HSI, YVbCr, LAB and LUV) and thus provided new information. The color spaces that were linear combinations of RGB (XYZ and YIQ) were, however, still expected to provide complementary information. The reason for believing so was based on previous work from Harrabi et al. and Mignotte who have had successful results using the chosen color spaces [20][22].

Only one channel was used from each color space. The channels were chosen by the amount of information that they gave about the image. In most color spaces the channel that described the intensity was the most informative. This was the case in HSI, YCbCr, LUV and YIQ.

3.3.3.3 Dempster Shafer Theory

Dempster Shafer Theory is also known as Evidence Theory. This is because evidence from different sources are used and evaluated if they are reliable or not. In this context the evidence is represented by the initial Otsu segmentations of the color channels in respective color spaces. As previously mentioned the evidence needs to be evaluated on how probable it is. The evidence provides a hypothesis, for instance that the pixel is foreground, and the reliability of the hypothesis is evaluated with a belief function. The belief functions from different sources are then combined to provide a final decision of what the most probable answer is. The theory behind belief functions was introduced by Arthur P. Dempster in 1967 [23]. The work was then further developed by Glenn Shafer in 1977 resulting in the evidence theory [24].

3.3.3.4 Power set

The power set includes all possible states of a system and is denoted as: 2^x

If X is a set: (a,b) the set includes following subsets:
-a (foreground)
-b (background)
-(a,b) (both foreground and background)
- the empty set $\emptyset$

The power set 2^x includes all the different subsets:

$$2^x = \{\emptyset, a, b, (a, b)\} \tag{3.23}$$

Since the set X consists of two components (foreground and background) the power set includes 2^2 different subsets of possible states.

3.3.3.5 Mass function

The belief mass function, or simply the mass function, assigns a belief mass to every subset i.e. how probable each subset is on a level from 0 to 1:

$m : 2^x \rightarrow [0, 1]$
One property of the mass function is that the empty set $\emptyset$ has a belief of 0. This means that there has to be either background, foreground or a combination in every image. This requirement is obviously fulfilled in this case since every part of the image is either foreground or background.

The second requirement is that when adding the mass functions of every subset together it has to equal to 1:

$$\sum_{A \subseteq X} m(A) = 1. \tag{3.24}$$

One can interpret this as the total belief of 1 is distributed over the evidence available. The mass function can be chosen in various ways depending on the scope of the problem. The following function for calculating the mass function was used in this application:

$$m^q_{x,y}(C_i) = \frac{1}{\sigma_i \sqrt{2\pi}} e^{\frac{-(g^q_{x,y} - \mu_i)^2}{2\sigma_i^2}}. \tag{3.25}$$

The function was proposed by Chaabane et. al with an assumption that the pixels are Gaussian distributed [25]. C_i is the class from the inital segmentation (foreground and background), q is the number of the feature i.e. the number of the color space used, μ_i is the mean value of the whole class, σ_i^2 is the variance of the class and $g_{x,y}$ is the value of the pixel at position (x,y).

In addition, the neighbouring pixels were taken into account by calculating the mass function presented in equation 3.25 within a quadratic window with a length of three pixels:

$$m^q_{x,y}(C_i) = \frac{1}{t^2} \sum_{x'=x-(t-1)/2}^{x+(t-1)/2} \sum_{y'=y-(t-1)/2}^{y+(t-1)/2} m^q_{x',y'}. \tag{3.26}$$

In equation 3.26 is t the length of the window, $m_{x'y'}$ is the mass function calculated within a window using equation 3.25 and (x, y) is the position of the current pixel.

The size of the window is a trade off between noise and preserving details. If the window is too large information about the details are lost, for instance where the border is located. However, a larger window is more stable to noisy pixels. Since the bacteria are very small, only a few pixels wide, a

small window with the length of 3 pixels was chosen to be able to accurately capture the details.

Figure 3.9 shows an example of a mass function that was calculated using this method with evidence from the LAB color space, using the A-channel.

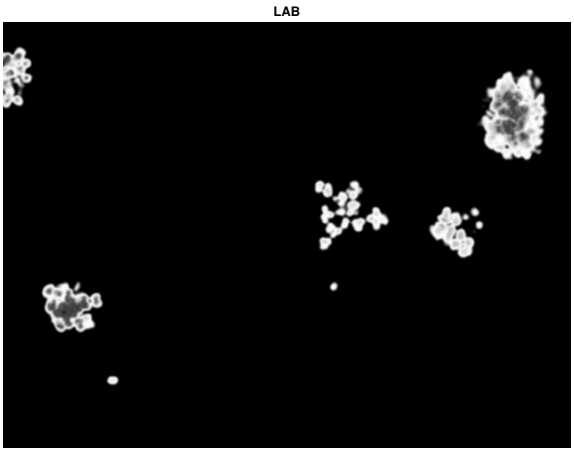


Figure 3.9: Mass function for class 1 calculated from the LAB-colorspace, using the A-channel.

A high value of the mass function (bright pixel) means that the probability of a pixel belonging to a certain class is high. The mass function was in this case calculated for three different hypothesis; the pixel belongs to class 1 (foreground), the pixel belongs to class 2 (background) or the pixel is either class 1 or class 2. The last category has in general a high value at the borders of the bacteria since it is hard to tell where to separate the bacteria from the background. The mass function calculated for the three different hypothesis can be seen in Figure 3.10.

3.3.3.6 DS rule of combination

As previously mentioned, the idea was to merge the information from the different sources to obtain a final answer. The available information now consisted of mass functions from all the different color spaces. To enable a combination of the mass functions from the different sources, DS rule of combination was used. DS rule of combination compile a shared belief from

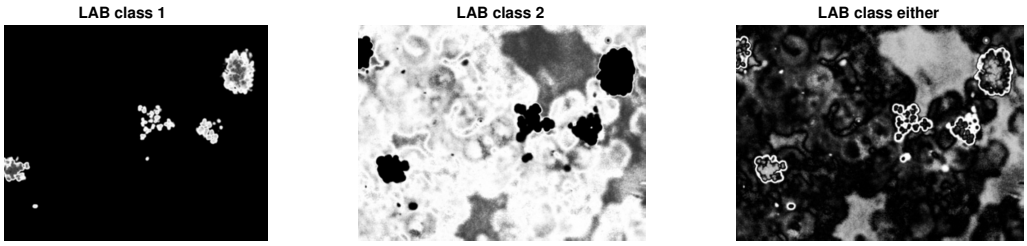


Figure 3.10: Mass function for class 1, class 2 and class 1 or class 2.

different sources and ignores conflicting beliefs. In practice this is done by orthogonally adding the mass functions from the different sources [20]:

$$m_{x,y}(C_i) = m^1_{x,y}(C_i) \oplus m^2_{x,y}(C_i) \oplus ... \oplus m^q_{x,y}(C_i). \tag{3.27}$$

Where C_i is the class and q is the number of the color space. The summation is associative and commutative [20]. In practice, two mass functions are combined in the following way:

$$m^{1,2}_{x,y}(C_1) = \frac{1}{1-K}(m^1_{x,y}(C_1)m^2_{x,y}(C_1)+m^1_{x,y}C(_1)m^2_{x,y}(C_1 \cup C_2)+m^1_{x,y}(C_1 \cup C_2)m^2_{x,y}(C_1)), \tag{3.28}$$

$$m^{1,2}_{x,y}(C_2) = \frac{1}{1-K}(m^1_{x,y}(C_2)m^2_{x,y}(C_2)+m^1_{x,y}(C_2)m^2_{x,y}(C_1 \cup C_2)+m^1_{x,y}(C_1 \cup C_2)\vec{m}^2_{x,y}(C_2)), \tag{3.29}$$

$$m^{1,2}_{x,y}(C_1 \cup C_2) = \frac{1}{1-K}(m^1_{x,y}(C_1 \cup C_2)m^2_{x,y}(C_1 \cup C_2)), \tag{3.30}$$

Where K, $m^1_{x,y}(C_1 \cup C_2)$ and $m^2_{x,y}(C_1 \cup C_2)$ are calculated in the following way:

$$K = m^1_{x,y}(C_2)m^2_{x,y}(C_1) + m^1_{x,y}(C_1)m^2_{x,y}(C_2), \tag{3.31}$$

$$m^1_{x,y}(C_1 \cup C_2) = 1 - m^1_{x,y}(C_1) - m^1_{x,y}(C_2), \tag{3.32}$$

$$m^2_{x,y}(C_1 \cup C_2) = 1 - m^2_{x,y}(C_1) - m^2_{x,y}(C_2). \tag{3.33}$$

K is a normalizing factor and measures the amount of conflict between the two sources [20]. Since the summation is associative one can add the third mass function to the combined mass function from source 1 and 2 ($m^{1,2}_{x,y}$). The combination of mass functions can continue by adding more mass functions as described above.

As stated before, DS rule of combination ignores beliefs that are conflicting. This property requires that the belief from the different sources are somewhat conformed. As a consequence of this property it is crucial that the different color channels are chosen adequately.

3.3.3.7 Choosing color channels

Several papers (Chabaane et.al, Harrabi et.al) address the need of choosing the color channels appropriately so that the evidence contains as much valuable information as possible [20][26]. To measure the sensitivity of the color channel Chabaane et.al and Harrabi et.al made an initial segmentation using every color channel in every color space. The segmentation was then compared to ground truth. The color channel in each respective color space that had produced the most accurate segmentation was chosen as a source in the DS method for segmentation. However, these papers had access to ground truth of foreground background belonging for each pixel, which was not available in this thesis. Instead of using ground truth a visual assessment was done to determine the most sensitive color channel in every color space. The user determined which and how many color channels that should be included as evidence to enable a satisfactory segmentation.

3.3.3.8 Segmentation of both Gram negative and Gram positive bacteria

DS-segementation supports multi-level thresholding and therefore an intial Otsu segmentation using three classes (two for bacteria and one for background) was tested for images with both Gram positive and Gram negative bacteria [20]. The initial segmentation using Otsu failed however to determine the two bacteria as two different classes. Instead it concluded that the bacteria was one class and that part of the background (such as blood cells) was the second class. Consequently, using multi-level thresholding was ruled out as an option due to the interfering background.

3.4 Detection and Classification

Information can be retrieved from the image with or without segmentation. The bacteria detection section below uses the original image. The other methods use the objects from the segmented image.

3.4.1 Bacteria detection

In order to retrieve as much information as possible from each image it is of interest to detect all bacteria and their location. To find the positions of the bacteria the image had to undergo a few transformations. First the image was dilated with a black top hat filter of size 6 x 6 pixels. The size of the filter was chosen based on the size of cocci bacteria in the images. The concept of dilation is explained in the paragraph below. A black top hat filter has zeros in the middle and ones in the corners. Using this filter the surroundings of the bacteria were dilated instead of the middle. The result was that the peak in the middle of the bacteria decreased in intensity in the dilated image, an example can be seen in Figure 3.11. Secondly, the dilated image was compared to the original image shown in Figure 3.12. All locations where the original image had a higher intensity were saved. The resulting image was dark with bright peaks where the local maxima were located.

Dilation is an example of a simple morphological operation that is related to the form and structure of an image. It have many different useful areas, but most often it is used to remove imperfections in the image created by for example the camera or thresholding [27]. Dilating depend on a structuring element that is used to transform the image. A structuring element is a real-valued 2D function. $H(i, j) \in \mathrm{R}$ for $(i, j) \in Z^2$ [28]. The size and shape of the structuring element changes the amount and kind of effect the dilation has on the original image. In all morphological operations the value of a pixel in the output image is based on a comparison of the same pixel in the original image with its neighbours [29]. How many neighbours and in which direction the neighbours are chosen is based on the structuring element. Dilation is done by finding the maximum of the values in the structuring element H added to the sub-image I, that is a part of the original image.

$$(I \oplus H)(u, v) = \max_{(i,j)\in H} I(u + i, v + j) + H(i, j). \tag{3.34}$$

Basically, dilation thickens the objects in an image by adding pixels around the boundaries. Dilation is closely related to the method erosion which instead makes the objects thinner by removing pixels by the borders. In both cases the result is an image with the same size as the original image. The procedure is used for gray scale images.

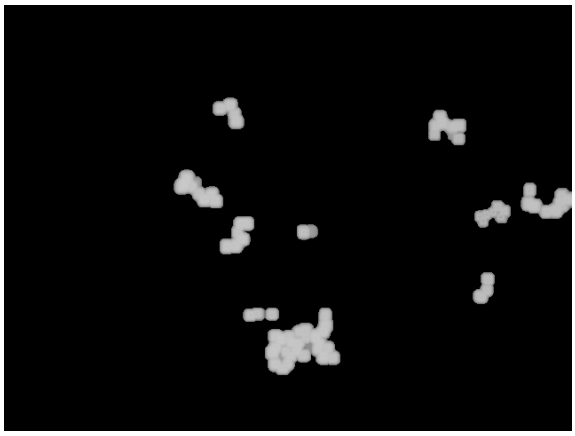


Figure 3.11: An image of cocci that has been dilated with a top-hat filter.

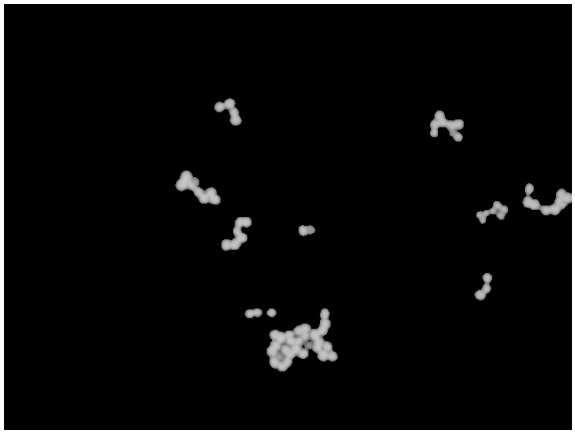


Figure 3.12: The original gray scale image of cocci.

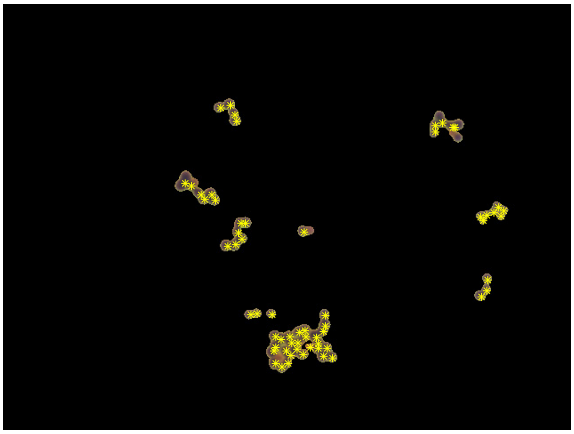


Figure 3.13: The bacteria have been detected in the image. The position where a bacterium is detected is marked in yellow.

Some local maxima are false and need to be detected and deleted, though these maxima are rare. To do so the distance to a neighbouring maximum is not allowed to be closer than a certain threshold. The threshold was chosen to be the same amount of pixels as the diameter of a small cocci. Not only is the location of the bacteria acquired with this method, the number of bacteria in an image can also be counted. In Figure 3.13 the detected bacteria using this method are visualized.

3.4.2 Label connected components

Objects in an image, such as bacteria, are assumed to be one connected object if they are constituted of pixels that are neighbours which are similar to each other. In order to keep track of the features of different objects in an image it is of interest to label the objects with a labeling algorithm. The algorithm is an already implemented function in MATLAB, `bwlabel`, which was used frequently in the execution of the other methods in this section. The input in the MATLAB function is a binary image, and hence the similarity mentioned above constitutes of pixels which has values of one.

The labeling algorithm used in this thesis was presented by Haralick and Shapiro [30]. In the binary images the foreground has a value of 1 and the background a value of 0. If there are eight or more connected pixels with a

value of 1 the object gets a label. If there are for example 10 bacteria in an image and the area of each bacterium is larger than 8 pixels in the image, all the bacteria will get a label between 1 and 10. The labeling starts with the object that are furthest to the left which gets label 1. The algorithm then continues labeling connected objects from left to right in the image [30].

3.4.3 Fourier descriptors

In image analysis Fourier descriptors are used to describe the contour of an object. In this particular case, Fourier descriptors were used to enable a separation of objects with smooth contours, such as single cocci and bacilli but also larger structures such as red blood cells, from clusters with more irregular boundaries. The Fourier descriptors were thus used as a feature to classify different types of bacteria and as a preprocessing tool to exclude red blood cells from the image.

Zhang at al. have proposed a method for shape retrieval in medical images using Fourier descriptors, the retrieved features are invariant to rotation and translation [31]. The following equations for computing Fourier descriptors were based on Zhang et al.'s method.

The boundaries of the objects were found and the boundary pixels were described as:

$$P = \{(x_t, y_t) | t \in [1, N]\}. \tag{3.35}$$

Where P is the pixel set of the boundary pixels, x_t and y_t are the boundary pixel coordinates and N is the number of boundary pixels. The number of boundary pixels, N, has a correlation with the size of the object. This information will be used later.

Once the boundary pixels have been found, the euclidean distances, r_t, from the centroid and the boundary pixels of each object were calculated. The centroid coordinates of the object were found by taking the mean value of the boundary pixels:

$$x_c = \frac{1}{N} \sum_{t=0}^{N-1} x_t, \tag{3.36}$$

$$y_c = \frac{1}{N} \sum_{t=0}^{N-1} y_t. \tag{3.37}$$

The calculation of the euclidian distances was thereafter done in the following way:

$$r_t = ([x_t - x_c]^2 + [y_t - y_c]^2)^{1/2}, t = 0, 1, ..., N - 1. \tag{3.38}$$

The discrete Fourier transform of r_t was then calculated. The Fourier transformed coefficients of r_t can be denoted as:

$$a_n = \frac{1}{N} \sum_{t=0}^{N-1} r_t e^{\frac{-2j\pi nt}{N}}, n = 0, 1, ..., N-1. \tag{3.39}$$

In equation 3.39 a_n is the discrete Fourier transform of of r_t. When taking the Fourier transform information about the magnitude and the phase are retrieved, in this case only the magnitude was of interest and consequently the phase was ignored. To create shape features every Fourier coefficient was standardized with the first Fourier coefficient: $b_n = |a_n/a_0|$ This gave a measurement of how complex the boundary of the object was. If the object was round the coefficients were small. A more complex boundary requires larger Fourier descriptors to be fully described. The reason for this is because an irregular boundary needs more frequency components to be satisfactionally represented by Fourier coefficients. Descriptors that are too small are considered being unimportant for describing the shape.

3.4.4 Remove red blood cells and debris

In order to be able to analyze the images properly, objects that are not bacteria need to be removed. In order to identify and separate disturbing objects from bacteria, all the objects in the image were labeled as described in section 3.4.2. The area of the labeled objects was then calculated. Objects whose area was smaller than a single bacterium were assumed to be noise and were therefore removed. Objects at the border of the image were also removed since only a part of the objects were captured.

A common disturbing object in the images are red blood cells. The blood cells are oftentimes absorbing the stain resulting in a color similar to Gram negative bacteria. Consequently, they can not be separated based on their color. Instead, they need to be separated with regard to their size and shape. Red blood cells are larger than bacteria and have most often a round shape.

To find a feature describing the shape, Fourier descriptors were used (described in section 3.4.3). A round object is characterized by its very small Fourier coefficients. A more irregular object give rise to larger Fourier descriptors. Since the red blood cells are round, they were identified by their small Fourier descriptors and long boundary.

An example of the preprocessing using the Fourier descriptors can be seen in Figure 3.15a. Figure 3.15a shows an original segmentation for Gram

negative bacteria. Figure 3.15b shows the segmentation after the red blood cells had been removed by identifying the red blood cells with their Fourier descriptors and the size. However, as can also be seen in Figure 3.15b there are some structures left that did not fulfill the criteria. It can be seen in Figure 3.14 that these structures are in fact red blood cells, though a bit damaged by the bacteria that are lying on top of them. It can also be seen that the segmentation in Figure 3.15b compared to Figure 3.14 has left out more details and hence the round shapes are not complete.

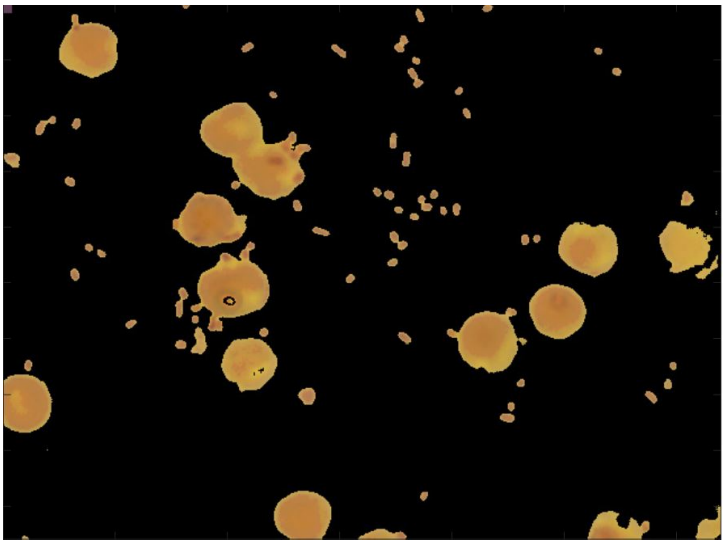


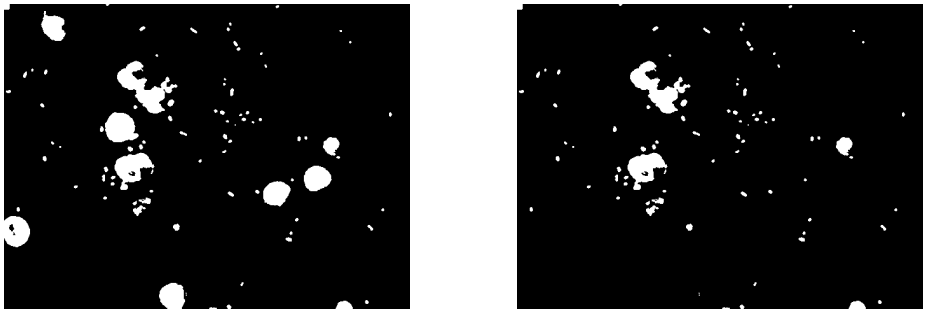
Figure 3.14: An image of the segmentation of the negative bacilli.

Clearly, only using Fourier descriptors was not enough to remove all the blood cells. To find red blood cells whose shape have been altered by for instance overlapping bacteria, template matching was used.

3.4.4.1 Template Matching

The template matching algorithm finds the matching part of the image by calculating the sum of the absolute values of the differences between the pixels in the templates and in the image. This algorithm is an already existing template matching system in MATLAB, `step`, and will be described in greater detail below [32].

The region that gets the lowest score with one of the templates is considered to be the matching area. The sum of absolute differences (SAD) is



(a) A segmentation with red blood cells present.

(b) Same segmentation after the red blood cells have been removed.

Figure 3.15: Removal of red blood cells.

used as a score of how good the match is and is calculated in the following way [33]:

$$d_1(I_j, T) = \sum_{i=1}^{n} |I_{i,j} - T_i|. \tag{3.40}$$

Where I is the image, j is the place in the image the template is compared to and T is the template. A standard red blood cell was used as the template T. Matches that were good enough, i.e. matches that have an SAD-value below a certain value, were removed. This value was found by doing several template matches in the same image. When the template started to match with objects that were not red blood cells the SAD-value was too high. A SAD-value below 5500 was considered to be a good threshold to distinguish correct matches from matches with objects that were not red blood cells.

3.4.5 Classify bacteria as Gram positive or Gram negative

To classify a bacterium as Gram negative or Gram positive the color of the bacterium needed to be found. The bacterium can have a color that varies somewhat within the shape, both in intensity and shade as can be seen in 3.16. The point where the color intensity has its peak was therefore chosen to represent the color of the whole bacterium. This point was found using the top-hat filter described in the Bacteria detection section 3.4.1.

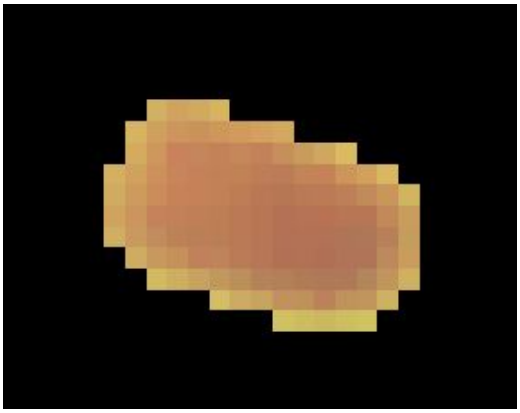


Figure 3.16: Gram negative bacilli where the intensity gradient can be seen.

The peak point created a vector with the three colors in RGB which was then compared to the two standard stains used to color the bacteria in the stain normalization function, $\boldsymbol{s_1^*}$ and $\boldsymbol{s_2^*}$. The distance between the color vector of the bacterium and the two stain vectors were calculated according to equation 3.6 and resulted in two θ answers. The bacterium was classified as Gram negative if the θ created with the Gram negative stain vector was smaller than the θ created with the Gram positive stain vector. The bacterium was evidently classified as Gram positive if the Gram positive θ was smaller than the Gram negative θ.

3.4.6 Separating bacteria growing in groups, pairs and individually

Large groups of bacteria were easily classified based on their area. In the same way it would be easy to classify single bacteria upon their size since they are very small. However, the size of a single bacterium vary and especially between cocci and bacilli. Hypothetically bacteria in pairs should have a larger area than individual bacteria. The cocci that grow in pairs however are easily confused with singles of bacilli since their shape and size can be very similar, see Figure 3.17a and 3.17b.

(a) Cocci growing in pairs. (b) Bacilli growing individually.

Figure 3.17: Bacteria in pairs compared with bacilli.

3.4.6.1 Distinguishing pairs of cocci from single bacteria using top hat filter

A way to separate pairs from singles would be to check if there are two regions with higher color intensity within the object, indicating a pair. To find the color intensity peaks, a black top hat filter was used as described in the Bacteria Detection section 3.4.1. If the top hat filter found two peaks, it was classified as a pair. If the top hat filter found only one peak it was classified a single bacterium.

The method works decently but can sometimes detect two peaks in a bacillus since the color intensity can vary. As a result, there was a need for a complementing feature separating singles from pairs. The complementing feature will be explained below in the section Convex hull.

3.4.6.2 Convex hull

In a pair constellation, there is a concavity where the two bacteria meet. A way to measure how big the concavity is, is to use the convex hull and count the number of zeros that are included in the convex hull. If it is a pair, the number of zeros will be larger than the number of zeros for a bacillus. In Figure 3.18 a pair of cocci are shown in white. The convex hull of the pair structure is displayed in red. The convex hull algorithm is an already implemented function in MATLAB, `convhull`.

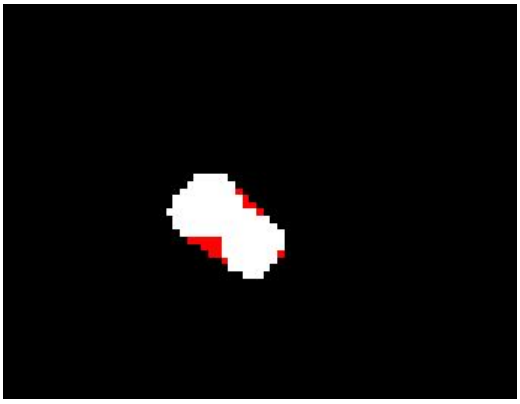


Figure 3.18: A binary image of two cocci lying above its convex hull in red.

Let S be a subset of the plane. S is called convex if and only if for any pair of points in the subset: $p, q \in S$, the line between the points $\overline{pq}$ is completely contained in S. The convex hull of S is defined as the smallest convex set that contains S.

A way to visualize the convex hull is to imagine having a set of points P and putting a rubber band around the points so that every point is surrounded by the rubber band. The rubber band will minimize its length and the area enclosed by the band will be the convex hull of P [34].

3.4.7 Classify larger structures of bacteria as chains or clusters

As mentioned before, different types of bacteria grow in different ways. Some grow as singles or pairs and some grow in larger structures such as clusters or chains. To be able to classify these bacteria it is important to separate cluster and chain formations from each other. The criteria for an object to be considered being a chain or a cluster was that the object included at least three bacteria.

The method chosen for the separation between clusters and chains was based on a skeleton function. For each object a skeleton was created which was equidistant to the boundary of the object. Kong and Rosenfeld describe the skeleton as being a stick-like representation of the object which gives information about shape properties such as elongation and width. The skeleton is placed on the medial region of the object [35]. An example of how an object's skeleton can look is shown in Figures 3.19b and 3.20b.

Skeletonization can be done in different ways, however this method was based on the concept of maximal disks. According to Wu, Merchant and Castleman skeletons have been successfully used for feature extraction to enable classification in microscopy images. The method of maximal disks looks for the largest disk possible that is still lying within the object. A disk is a round object. The centers of the maximal disks will create the medial axis of the image [36]. The skeleton function is a part of the `bwmorph` operations already implemented in MATLAB.

The final skeleton is created from Lantuejoul's formula:

$$Skel(S) = \bigcup_{n=0}^{\infty} (S \ominus nB) - [S \ominus nB \circ B]. \tag{3.41}$$

The skeleton Skel is the union of the skeleton subsets $Skel(S; n)$

$$Skel(S; n) = (S \ominus nB) - [(S \ominus nB) \circ B], \tag{3.42}$$

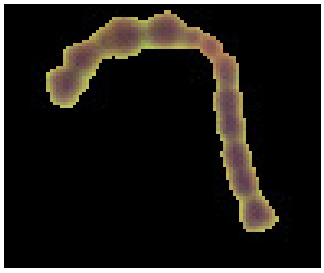
where nB is defined by

$$nB = B \oplus B \oplus ... \oplus B. \tag{3.43}$$

B is the structuring element, in this case a disk. nB is the iterated dilation of the disk used to find the maximal disk where the disk can no longer become larger without going outside the object. Hence, n is the number of times that a particular disk was dilated.

To create a 'cluster or chain' measure the shortest distance from each edge point to the skeleton was found. This gave a measure of how skinny the object was. The mean of the distances for a certain object was found and then divided by the major axis. Dividing by the major axis makes the measure more neutral to the size of the object. A visual assessment was made to find a threshold for when an object should be classified as cluster or chain. Clusters have a larger mean distance divided by major axis fraction than chains.

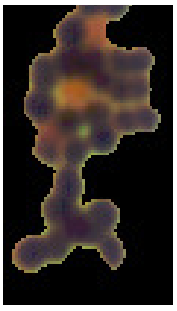
The difference between a chain and a cluster and their skeletons can be seen in Figures 3.19a, 3.19b, 3.20a and 3.20b. The skeleton in the cluster has branches whereas the chain has not. It is also clear from the figures that the distance from the border to the skeleton is larger when the bacteria are in a cluster formation than when the object has a small width like chains.



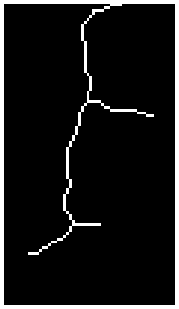
(a) Bacteria growing in a chain formation.

(b) Skeleton of the chain formation.

Figure 3.19: Chain formation



(a) Segmentation of bacteria growing in a cluster formation.



(b) Skeleton of the cluster formation.

Figure 3.20: Cluster formation.

3.4.8 Classify bacteria as cocci or bacilli

A straightforward way to discriminate cocci from bacilli is to compare the major axis to the minor axis. For cocci the major axis and the minor axis are approximately the same due to their round shape. For bacilli the major axis is longer since the bacterium is rod-shaped. Consequently, if the bacteria were growing individually the ratio between the major and minor axis was used to classify the bacteria as cocci or bacilli. The bacteria were first labeled (see section 3.4.2) in an image and then the ratio between major and minor axis was calculated. If the ratio between minor and major axis was larger than a certain threshold the bacterium was considered to be a coccus. If the

ratio was smaller than the threshold the bacterium was classified as bacilli. The threshold for the ratio between minor and major axis was chosen to be 0.65 since it provided an accurate distinction between cocci and bacilli. The threshold was determined by investigating several images of both cocci and bacilli.

It was however more complicated to classify bacteria as cocci or bacilli if they did not grow individually. Two different approaches were implemented to determine the shape of the bacteria. Both methods aim to overcome the problem with bacteria lying in pairs, chains or clusters.

3.4.8.1 Cluster segmentation using watershed

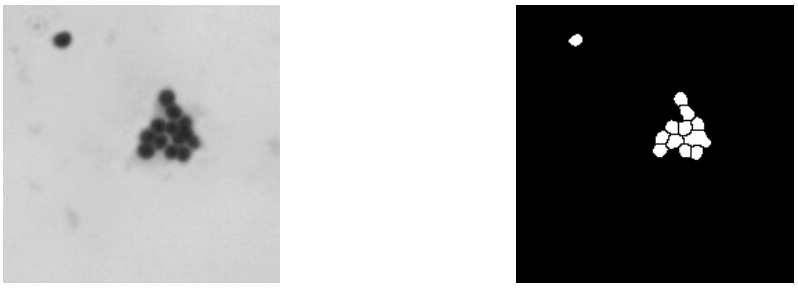
To enable a comparison of the major and minor axis for each bacterium they need to be segmented individually. To achieve an individual segmentation of bacteria in a cluster or a chain, a segmentation method using watershed was tested. The watershed method used in this thesis is based on the algorithm developed by F. Meyer [37].

3.4.8.2 Watershed by flooding

Watershed is a technique aiming to find the border between adjecent objects, in this case the aim was to find the border between adjacent bacteria. In Meyer's flooding algorithm the image is modeled as a topographic surface where the light pixels are high elevations and the dark pixels are low elevations. There are different ways to implement watershed and this particular algorithm, that was already implemented in MATLAB, `watershed`, uses watershed by flooding. Metaphorically, the darkest pixels in each region are the regional minima where the flooding starts. Each different minimum are flooded with water and when the water from different sources meet they build a barrier which makes up the watershed lines. These watershed lines are supposed to emerge at the edges of an object in an image.

F. Meyer's algorithm was used on gray scale images. The algorithm uses selected starting points for the flooding. In this case the top hat filter explained in the Bacteria Detection section 3.4.1 was used to find starting points, that are minima in a specific region. These starting points are located at the middle of the bacterium.

The flooding algorithm aiming to find the watershed lines uses a priority queue and is gradually labeling the pixels based on their gray scale value [38]. The labeling starts in the middle of the bacterium and continues labeling



(a) A bacteria cluster with cocci. (b) Segmented cocci using watershed.

Figure 3.21: Watershed segmentation on cocci cluster.

until it meets another label that has started from another bacterium. When this happens a border is constructed between the bacteria.

Figure 3.21a shows an example of how a cluster of cocci can look like in gray scale. In Figure 3.21b is the result of applying a watershed algorithm to separate the cluster, is shown.

3.4.8.3 Template matching to find cocci or bacilli

Another way to find different shapes in an image is to use template matching. The method was the same as described in the section 3.4.4 but other templates were used for this purpose. Two different groups of templates were used, one group aiming to detect cocci and one group for bacilli detection. The cocci templates were constructed from a representable cocci from one of the images and then scaled to three different sizes. The bacilli templates were made from an adequately shaped bacilli and then rotated in 18 different angles. An example of bacilli templates can be seen in 3.22.

The template matching was then performed on the bacteria growing in cluster or chains to see whether it was a bacilli or cocci template that had the best matching score. If a template of cocci had the best score the chain or cluster was considered to be constitued of cocci and vice versa for bacilli.

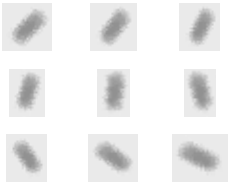


Figure 3.22: Bacilli templates rotated in different angles.

3.5 Overall System

The aim of the thesis was to be able to correctly detect bacteria in an image and then classify bacteria as:

- Gram negative or Gram positive
- Cocci or bacilli
- Singles, pairs, chains or clusters

In the overall system have different components been put together to achieve the above requirements. In Figure 3.23 a flowchart of the method can be seen.

First of all a segmentation was done using Dempster Shafer theory described in section 3.3.3. Thereafter objects that were not bacteria such as red blood cells were removed using the method in section 3.4.4. When only bacteria remained in the image the bacteria detection method in 3.4.1 using a top hat filter was executed. A distinction between singles, pairs and groups were then made as described in section 3.4.6.

New images were created, one image with only groups of bacteria, one with pairs and one with singles. The separation was done because different classification methods were used depending on if there were bacteria growing individually or not.

The bacteria gets classified as Gram negative or Gram positive using the method described in section 3.4.5. For the images with groups of bacteria, the constellation of the group gets classified as chain or cluster using the skeleton function in 3.4.7.

If there are single bacteria in an image they got classified as bacilli or cocci by calculating the ratio between the major and minor axis according to the method in section 3.4.6. If there are no single bacteria in the image,

the template matching function described in 3.4.8.3 was used. If there are an image with bacteria both growing individually and in groups or pairs an assumption was made that the pairs or groups of bacteria were constituted of the same type of bacteria as the individually growing bacteria. For example, if their was an image with clusters and individual bacteria and the individual bacteria got classified as cocci it was assumed that the clusters also consisted of cocci. This assumption was made because the template matching function was not as accurate as the classification method for the individually growing bacteria. Finally, the positions where bacteria were detected as well as the resulting classification was obtained.

As stated in the title of the thesis, the overall system was semi-automatic instead of automatic. This is due to the manual selections in DS-segmentation.

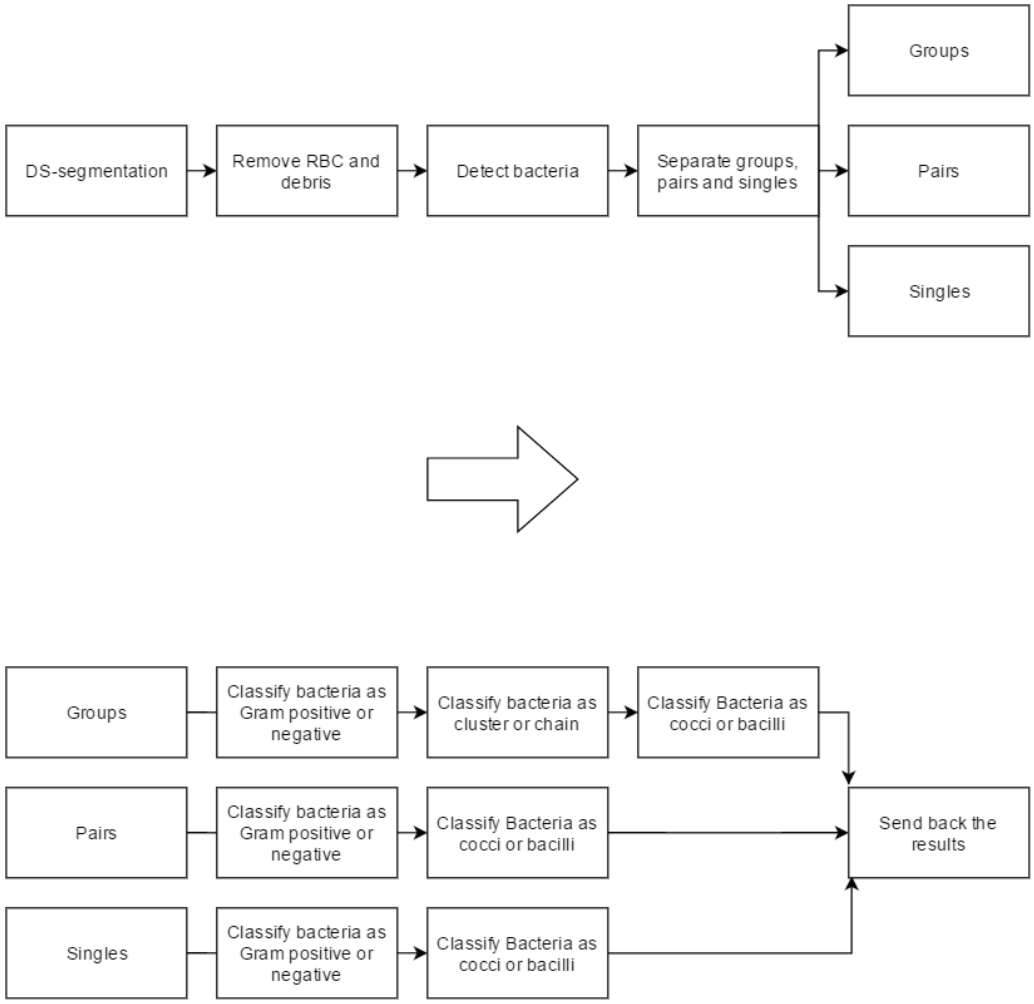


Figure 3.23: The flowchart of the overall method.

Chapter 4

Results

In this chapter the results from the methods are presented as well as the performance of the overall classification and detection system.

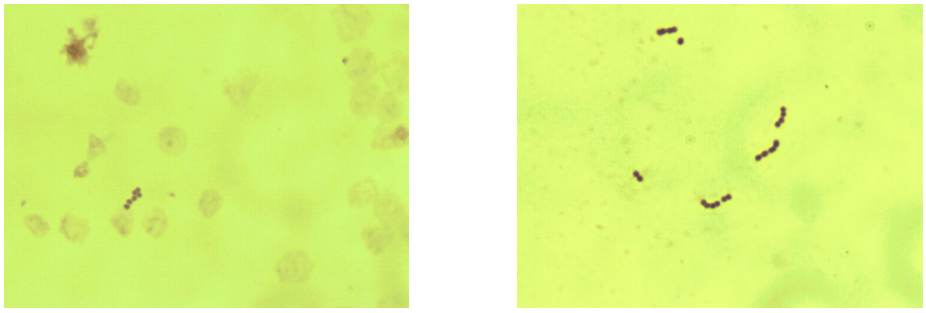
There were 90 different images available for testing from the new set of samples from the Microbiology laboratory in Lund that had not been used for training. The images show 9 different samples which included 10 different kinds of bacteria. An image analysis expert has created the ground truth by visually examining the images. The ground truth consisted of the location of the bacteria, if the bacteria were Gram negative or Gram positive, if the bacteria were in cluster or chain and if the bacteria were cocci or bacilli.

When evaluating Stain Normalization images of the samples from Copan and Calgary were used in addition to the samples from Lund.

The analysis of one image took about 2 minutes.

4.1 Stain Normalization

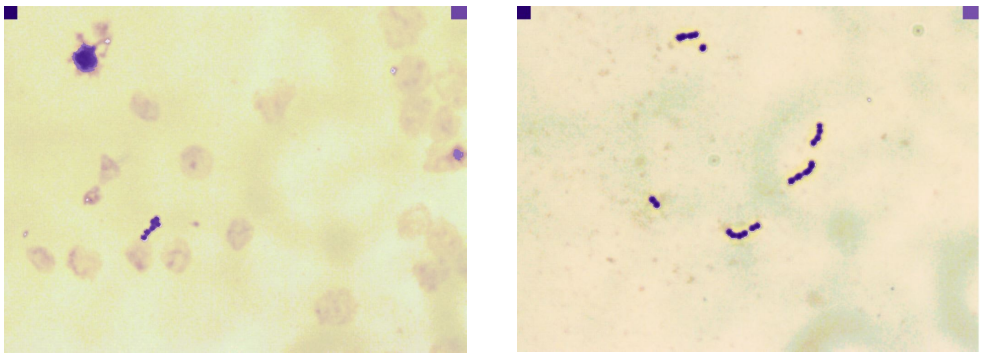
In this thesis there were samples from three different laboratories available, see chapter 2. The samples from two of the laboratories were four years old. The samples from the Microbiology laboratory in Lund was collected in the beginning of the thesis and were thus relatively newly prepared. The samples looked different depending on which of the laboratories they were collected from, see Figure 4.1a and 4.3b. As can be seen, both images contain Gram positive cocci but the older image is faded compared to the newer one. The fading could have been caused by exposure to light or differences in preparation. Figure 4.2a and 4.2b visualizes two examples of images that have been stain normalized where the stain vectors were chosen manually.



(a) An image of the old samples. (b) An image of the new samples.

Figure 4.1: Comparison between an image from the old samples and an image from the new samples.

The background has been altered to resemble how the samples look under a regular light microscope for the human eye. The alteration was done to make the experience as similar as possible to a microscope for a hypothetical laboratory professional.



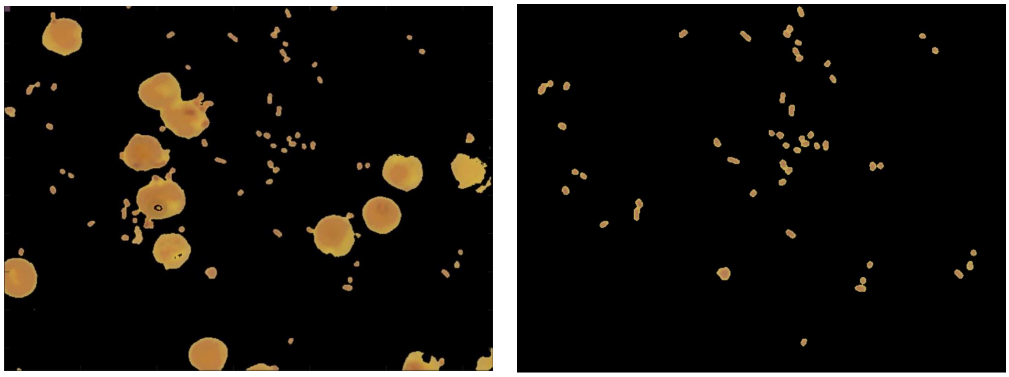
(a) Stain normalization of old image. (b) Stain normalization of new image.

Figure 4.2: A demonstration of stain normalized images.

4.2 Remove Red Blood Cells and Debris

The method that removes RBC and debris has not been evaluated separately with the validation data set. However, continuous visual assessments has

been made during the development of the algorithm. An example of how an image can look after the red blood cells have been removed using Fourier descriptors and template matching can be seen in Figure 4.3.



(a) Image before RBC removal.

(b) Image after RBC removal.

Figure 4.3: Before and after removal of Red Blood Cells.

4.3 DS-segmentation

In Figure 4.7 a comparison between images segmented with Otsu (using the MATLAB-function `graythresh`) and DS-segmentation can be seen.

When testing the results of DS-segmentation the segmented images received a grade on a scale from 1 to 3 based on a visual assessment. Grade 1 corresponds to an unsatisfactory segmentation where either parts of the bacteria have not been segmented or that objects which are not bacteria have been segmented. An image received a grade of 2 if the segmentation was decent. Grade two includes images where for example small non-bacterial objects have been segmented. Grade 3 corresponded to a perfect segmentation. In Figure 4.4, 4.5 and Figure 4.6 there are three segmented images, one of each grade. In Table 4.1 the results of the segmentation of the validation images are presented. The mean value for all images in a sample and the standard deviation (std) were calculated.

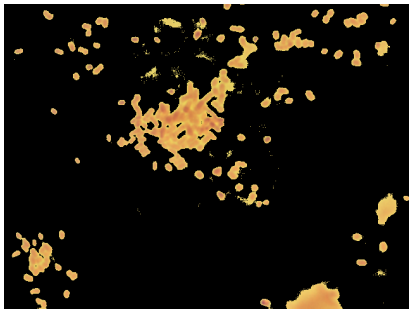
Figure 4.4: A segmented image with grade 1. This was a bad segmentation because objects that are not bacteria have also been segmented.

Figure 4.5: A segmented image with grade 2. It received grade 2 because of the small objects that have been segmented that are not bacteria.

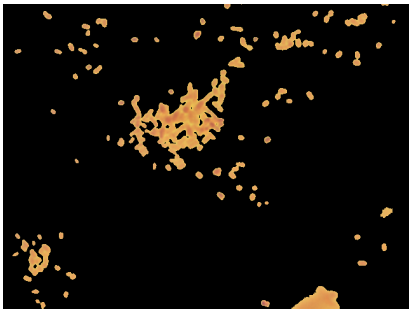
Figure 4.6: A segmented image with grade 3, the segmentation is perfect.

Table 4.1: Segmentation grade of DS-segmentation

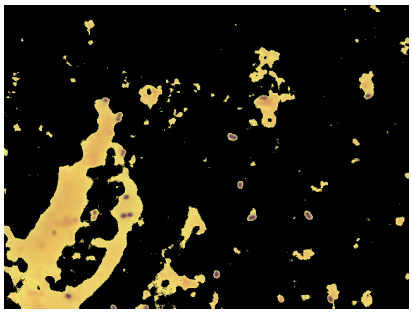
Name of bacteria	Type of bacteria	Segmentation grade	Std
E. coli	Gram negative bacilli	2.3	0.48
E. faecium + S. epidermidis	Gram positive cocci	2.7	0.48
K. pneumoniae	Gram negative bacilli	1.8	0.44
S. anginosus	Gram positive cocci	2.3	0.48
S. aureus	Gram positive cocci	2.1	0.88
S. aureus + S. pneumoniae	Gram positive cocci	2.1	0.78
S. epidermidis	Gram positive cocci	2.4	0.70
S. pneumoniae	Gram positive cocci	2.2	0.79
S. pyogenes	Gram positive cocci	2.9	0.32



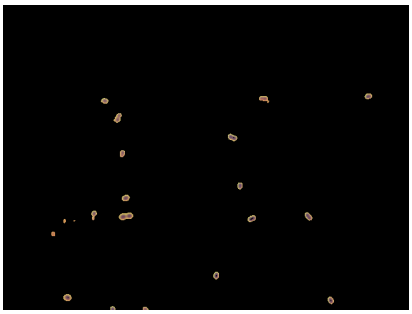
(a) Otsu segmentation.



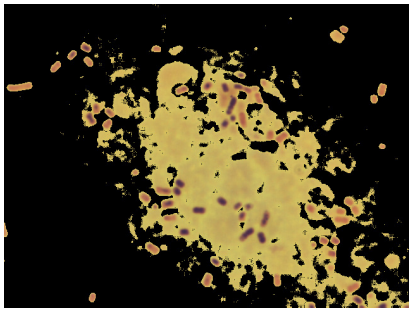
(b) DS-segmentation.



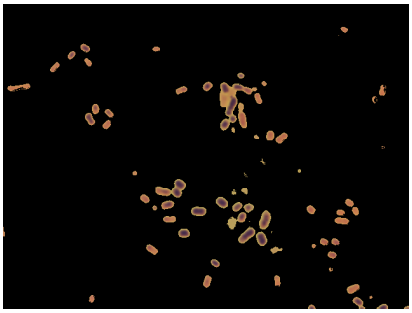
(c) Otsu segmentation.



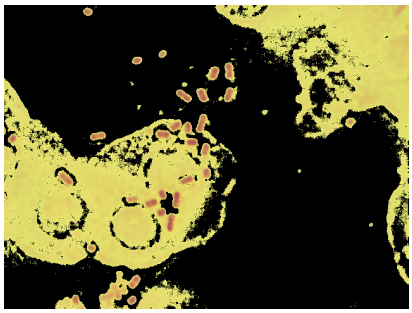
(d) DS-segmentation.



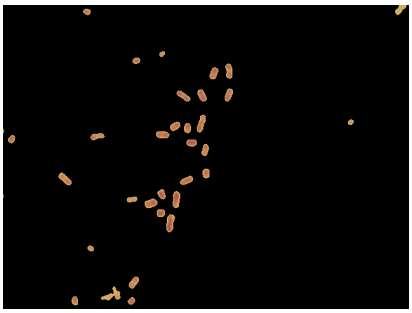
(e) Otsu segmentation.



(f) DS-segmentation.



(g) Otsu segmentation.



(h) DS-segmentation.

Figure 4.7: Comparison of images segmented with Otsu (to the left) and DS-segmentation (to the right).

4.4 Bacteria Detection

One of the aims was to detect the individual bacteria in the samples. How the bacteria are found is described in section 3.4.1. In Table 4.2 the results of the evaluation of the method is presented. The numbers of correctly found bacteria are presented as a percentage for each sample. The number of bacteria for the ten images in each sample was added together to a total percentage. A percentage of 100% means that all bacteria were correctly found.

Table 4.2: Percentage of correctly detected bacteria

Name of bacteria	Correctly detected	Type
E. coli	24.4%	bacilli
E. faecium + S. epidermidis	93.5%	cocci
K. pneumoniae	59.8%	bacilli
S. anginosus	91.3%	cocci
S. aureus	79.9%	cocci
S. aureus + S. pneumoniae	89.1%	cocci
S. epidermidis	92.9%	cocci
S. pneumoniae	89.0%	cocci
S. pyogenes	94.9%	cocci

4.5 Classify Bacteria as Gram Positive or Gram Negative

Another aim of the thesis was to be able to classify whether the bacteria were Gram positive or Gram negative, a description of the method can be found in section 3.4.5. The results from the algorithm and the ground truth for each sample can be seen in Table 4.3. GT stands for Ground Truth, A stands for Algorithm, GP stands for Gram Positive and GN stands for Gram Negative.

Table 4.3: Count of Gram positives and Gram negatives

Name of bacteria	GT GP	GT GN	A GP	A GN
E. coli	0	945	0	945
E. faecium + S. epidermidis	173	0	173	0
K. pneumoniae	27	342	29	340
S. anginosus	580	0	577	3
S. aureus	358	0	358	0
S. aureus + S. pneumoniae	360	5	352	13
S. epidermidis	230	2	230	2
S. pneumoniae	282	30	285	27
S. pyogenes	0	968	1	967

4.6 Classify Bacteria as Pairs or Singles

In Table 4.4 the method separating singles from pairs in section 3.4.6 was tested. The percentage of correctly detected singles and pairs were calculated, where 100% corresponds to that all pairs or singles have been correctly classified. CD stands for Correctly Detected.

Table 4.4: Percentage of correctly detected singles and pairs

Name of bacteria	CD singles	CD pairs
E. coli	79.4%	47.2%
E. faecium + S. epidermidis	81.7%	85.0%
K. pneumoniae	56.9%	13.1%
S. anginosus	77.1%	71.9%
S. aureus	67.0%	65.0%
S. aureus + S. pneumoniae	55.0%	66.7%
S. epidermidis	90.0%	75.0%
S. pneumoniae	71.5%	79.3%
S. pyogenes	100%	100%

4.7 Classify Bacteria as Chain or Cluster

In Table 4.5 the percentage of correctly detected and classified clusters and chains for each sample is presented. The method tested is described in section 3.4.7. CD stands for Correctly Detected.

Table 4.5: Percentage of correctly detected clusters and chains

Name of bacteria	CD clusters	CD chains
E. coli	46.5%	39.5%
E. faecium + S. epidermidis	58.3%	58.0%
K. pneumoniae	63.0%	5.6%
S. anginosus	70.0%	36.7%
S. aureus	72.8%	65.0%
S. aureus + S. pneumoniae	80.0%	83.0%
S. epidermidis	55.8%	31.7%
S. pneumoniae	79.3%	90.0%
S. pyogenes	80.0%	95.0%

4.8 Classify Bacteria as Cocci or Bacilli

There were several methods investigated to separate each individual bacterium as coccus or bacillus. If there were single bacteria in an image the ratio of major and minor axis was used to classify the bacteria as cocci or bacilli. If there were no single bacteria present other methods had to be used to identify their morphology. The two methods evaluated were cluster separation using watershed described in section 3.4.8.1 and template matching to find cocci or bacilli from section 3.4.8.3. For the final system a combination of methods was used.

To test how well the watershed method is able to distinguish between bacilli and cocci the ratio between the major and minor axis for each segmented bacteria in randomly selected images with clusters or chains was calculated. The mean value of the ratio for all bacteria in an image was then calculated as well as the standard deviation (std). The hypothesis was that a bacillus will have a higher ratio between the major and minor axis

than a coccus. Table 4.6 shows the result. Since this was an initial test whether or not the method was adequate the training data set images were used. In this set E. coli and S. pneumoniae was present in two samples which is why they are represented twice in Table 4.6. This applies to the template matching as well.

The template matching method was tested on three randomly selected images containing bacteria in chains or clusters from each of the 9 different samples, adding up to a total of 27 images. The method executed a template matching ten times in the image. When a region had been matched it was removed and the template matching continued with the rest of the image. The results can be seen in Table 4.7. The table shows the number of correctly matched bacilli or cocci as well as the total Sum of Absolute Differences (SAD) for all the 30 matches for every sample. SAD is a measurement of how good the match is, since it measures how much the template coincides with the matched region in the image.

If there were only clusters or chains in the image, the template matching function was used and not the watershed. The results from the bacilli and cocci classification method for the final system is found in Table 4.8.

Table 4.6: The ratio between major and minor axis for cluster separation using Watershed

Name of bacteria	Type	Mean ratio	Std
Escherichia coli 1	Bacilli	1.63	0.04
Escherichia coli 2	Bacilli	1.73	0.18
Klebsiella pneumoniae	Bacilli	1.57	0.08
Streptococcus anginosus	Cocci	1.40	0.07
Staphylococcus aureus	Cocci	1.40	0.14
Staphylococcus epidermis	Cocci	1.43	0.03
Streptococcus pneumoniae	Cocci	1.50	0.21
Streptococcus pyogenes	Cocci	1.54	0.17

Table 4.7: Number of correct matches and total SAD for Template matching

Name of bacteria	Type	Correct matches	Total SAD
Escherichia coli	Bacilli	19/30	4224
Escherichia coli	Bacilli	11/30	4487
Klebsiella pneumoniae	Bacilli	12/30	4358
Streptococcus anginosus	Cocci	30/30	3635
Staphylococcus aureus	Cocci	25/30	4489
Staphylococcus epidermis	Cocci	14/30	5408
Streptococcus pneumoniae	Cocci	29/30	3917
Streptococcus pneumoniae	Cocci	30/30	3308
Sterptococcus pyogenes	Cocci	29/30	4119

Table 4.8: Percentage of correctly detected Cocci or bacilli in final system

Name of bacteria	CD cocci or bacilli	Type
E. coli	30%	bacilli
E. faecium + S. epidermidis	70%	cocci
K. pneumoniae	50%	bacilli
S. anginosus	70%	cocci
S. aureus	100%	cocci
S. aureus + S. pneumoniae	60%	cocci
S. epidermidis	100%	cocci
S. pneumoniae	70%	cocci
S. pyogenes	100%	cocci

Chapter 5

Discussions, future development and conclusions

In this chapter the results of the implemented methods are discussed as well as suggestions for future developments and conclusions. It is important to have in mind that the methods were tested on quite few images. In order to get results that are more reliable more images would have to be used for testing. Moreover, the ground truth should ideally have been created by several microbiology experts.

However, the aim of the thesis was to investigate weather it was possible to detect and classify bacteria and the results provide an indication on what is possible to do.

5.1 Stain Normalization

By visually studying the stain normalized images, it was concluded that the automatic method for finding the stain vectors works well on images where both Gram negative and Gram positive stains are present. However, it does not work well when there is only one type of Gram stain present in the image. This is expected though since the method is developed for two stains, Gram negative and Gram positive. The method works decently when only Gram positive bacteria are present. This is probably because the Gram negative counter stain is absorbed to a certain extent by Gram positive bacteria and other substances such as white blood cells as well. Consequently, Gram negative stain is, to a certain degree, present in the majority of the images.

Stain normalization was not part of the final system. The reason for

that was because images that had been stain normalized sometimes got incorrectly stained which would then affect the rest of the processing. Moreover, the method was time consuming especially when the stain vectors were manually chosen. It is nevertheless an interesting method since it enables comparisons between samples whose staining properties varies.

5.2 Remove Red Blood Cells and Debris

As described in section 3.4.4 the red blood cells and other debris that are included in the segmentation are removed by two different methods, Fourier descriptors and template matching. The debris was only removed by the Fourier descriptors whilst the red blood cells were removed by both methods. The Fourier descriptors could not handle red blood cells that were too uneven in the edges, since they then obtained values that were too similar to bacteria clusters. However the overall shape of the bacteria were still different from the red blood cells, and therefore a template matching method was used as a complement.

The template matching worked well on the majority of the images. However, sometimes only parts of the red blood cells absorbed enough counter stain to be segmented, an example can be seen in Figure 5.1. These structures did not meet the requirements of either method to be identified as a red blood cell and could therefore not be removed. This effects both the results for Bacteria Detection, see sections 3.4.1 and 5.4, and the results in Classify Bacteria as Chain or Cluster, see sections 3.4.7 and 5.7.

A development of this function could be to use more templates of different common sizes and shapes of red blood cells that are partially segmented to be able to identify them.

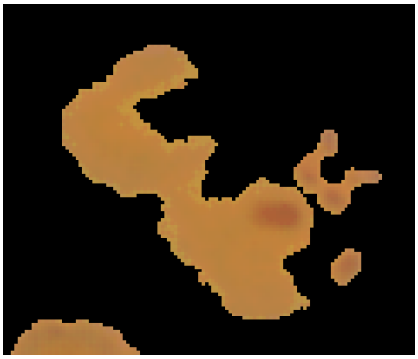


Figure 5.1: Red blood cells partially segmented.

5.3 DS-segmentation

The performance of the DS-segmentation can be seen in Table 4.1. The performance is somewhat better for Gram positive than Gram negative bacteria. The overall score for the Gram positive is 2.4 and the overall score for the Gram negative is 2.1. The quality of the segmentation varied among the images due to what kind of structures that were included in the images such as red blood cells. As a result, the standard deviations were overall quite high.

Nevertheless, DS-segmentation provides an improved segmentation compared to the standard Otsu segmentation. In Figure 4.7 the difference between Otsu and DS-segmentation are shown. As can be seen in the Figure, the Otsu segmentation oftentimes includes too much background. It is probably hard to automatically find a threshold separating bacteria from background since bacteria only make up a small part of the total image. Consequently, DS-segmentation was favorable for these images. Furthermore, it is crucial that the segmentation can be adjusted to be optimal for the current image which DS-segmentation provides, since the best segmentations from each of the color spaces can be chosen.

5.4 Bacteria Detection

As can be seen in Table 4.2 in section 4.4 the percentage of correctly identified bacteria varied between the samples. The biggest difference was between cocci and bacilli samples. Cocci samples had an overall percentage of 90.1 % whilst bacilli had an overall percentage of 42.1 %.

One explanation of the difference in performance is that the segmentation of the Gram negative stain is not as good as for the Gram positive and all of the bacilli in the test images were Gram negative. The algorithm then finds false positives in structures that are not bacteria.

Another problem with identifying individual bacilli is that they appear as if they have two kernels. An example can be seen in Figure 3.17b where the bacillus is compared to pairs of cocci. The algorithm counts these peaks as if they were two bacteria.

A conclusion from the results of the evaluation is that the current method works well for cocci detection. To improve the detection rate, a method that is better at detecting bacilli would have to be developed. One option could be to use a top-hat filter designed for bacilli. It would however be challenging to design such a filter since bacilli are not rotation invariant like cocci.

5.5 Classify as Gram Positive or Gram Negative

The classification of whether a bacterium was Gram positive or Gram negative received satisfactory results. As can be seen in Table 4.3 the results from the algorithm was 100 % correct in four out of nine samples. In the other five samples the algorithm was correct by 99.1 %. Hence, the results suggest that the method for identifying the stain was a good one. The drawback of this method is that it is based on the bacteria detection method, see section 3.4.1 for method and section 5.4 for results. The limitations of that method sequentially affects this classification.

5.6 Separating bacteria as groups, pairs and singles

5.6.1 Separate groups from pairs and singles

Groups of bacteria were separated from bacteria growing in pairs or individually solely based on the size of the area. Since the size of bacteria vary considerably among different species, size is not a reliable feature. This resulted in that some pairs of bacteria were mistakenly classified as groups of bacteria due to their large size. Moreover, some chains or clusters with

small bacteria were categorized as being a pair or individual bacteria. This was particularly a problem for the images with Streptococcus anginosus (S. anginosus) which are small and for the large Streptococcus epidermidis (S. epidermidis). In Table 4.5 one can see that both S. anginosis and S. epidermidis have a quite small detection rate of clusters and chains which in most cases was because the separation of groups from pairs and singles had been incorrectly done in the first place. If, for instance, a chain of small bacteria is classified as being a pair it will not be found when searching for chains or clusters thus causing a low detection rate.

In Figure 5.2 there is an example of two bacilli of large size which were falsely categorized as being chains.

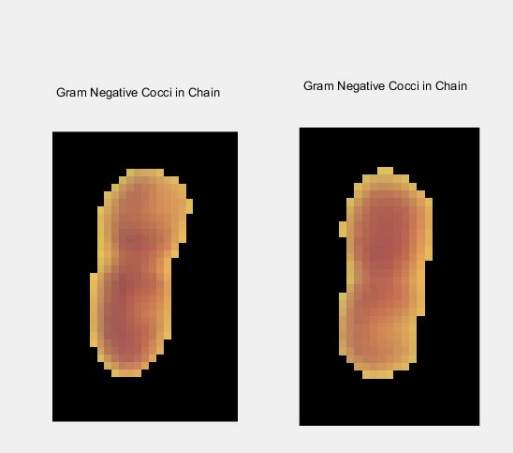


Figure 5.2: Two large bacilli wrongly classified as being chains.

5.6.2 Classify bacteria as pairs and singles

One source of error of classification of pairs and singles is the problems discussed in the above section. If the separation of groups from pairs and singles is incorrect, the classification of pairs and singles will also be incorrect. However, if the separation of groups from singles and pairs is correct there are still challenges remaining regarding classification of pairs or singles.

The method measuring the convex hull can be misleading if the bacterium is very small. As can be seen in Figure 5.3, the borders of the bacteria are not smooth due to the small size and limited resolution of the microscope.

There is thus many zeros included in the convex hull even though it is not a pair.

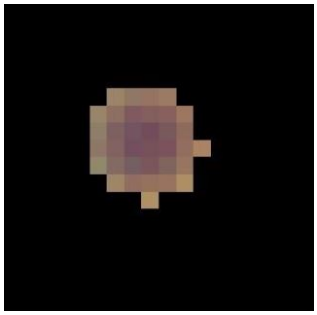


Figure 5.3: A small bacterium that mistakenly were classified as a pair.

The two Gram negative bacilli bacteria, Eschericia Coli (E. coli) and Klebsiella pneumoniae (K. pneumoniae), have a low detection rate of finding pairs as can be seen in Table 4.4. This is because the bacteria detection method works considerably worse for bacilli compared with cocci as described in section 5.4. K. pneumoniae has often two regions with a higher color intensity which Figure 5.4 visualizes. The bacteria detection function is then mistakenly classfying the bacilli as a pair.

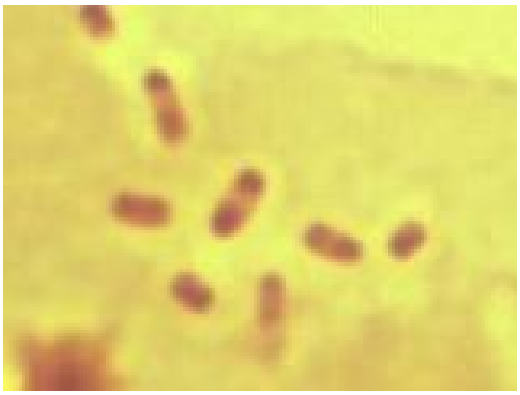


Figure 5.4: Bacilli with two regions of darker color.

5.7 Classify Groups as Cluster or Chain

The performance of the 'cluster or chain' classifier varies between different types of samples. The overall percentage of correctly classified clusters was 67.4 % and for chains it was 56.1 %. It was thus more difficult to classify chains than clusters. The reason for the difference in performance is the skeleton function used in the method, see section 3.4.7. The skeleton is sometimes not created all the way out on the chain. The distance between the skeleton and the border then becomes long and the classifier falsely classifies the chain as a cluster. The skeleton function used was MATLAB's own. A potential improvement of this classifier would be to create a new skeleton better adapted and more sensitive to these types of structures. Another development would be to include more classes. Some structures, for example with three bacteria, can be difficult to distinguish if they are chains or clusters and should therefore be a class of their own.

In some samples the classifier was 80 % correct for clusters and up to 95 % correct for chains. The clusters and chains in these samples are examples of typical clusters and chains with Gram positive cocci bacteria.

5.8 Classify bacteria as cocci or bacilli

5.8.1 Cluster segmentation using watershed

Because of the small differences in the ratio between major and minor axis between cocci and bacilli (see Table 4.6), watershed was not used in the final system. The reason for the small difference could be that the cluster segmentation have failed or because some bacilli has a round appearance. They can appear round for different reasons. Firstly, the bacilli captured might not be fully grown and therefore not as long. Secondly, their shape can be altered due to antibiotic treatment. Lastly, they can be located in the sample so that their short end is upwards facing the camera.

The cluster segmentation failed when bacteria were located too close to each other and there was no visual border between them. Consequently, the watershed algorithm was unable to separate those bacteria. Another source of error was that top-hat filter sometimes detected several minima in one bacterium.

5.8.2 Template matching

As can be seen in Table 4.7, the template matching is better at detecting cocci than bacilli. One reason that cocci are easier to match with could be because they are rotation invariant since they are round. Bacilli, on the other hand, can be rotated in many different angles and have various length and width. Furthermore, the color of bacilli are often weaker than the color of cocci since bacilli more often are Gram negative. This results in that other substances more often get falsely matched as bacilli. To improve the template matching more templates could have been used, especially bacilli templates. The drawback of using more templates is an increased execution time.

Template matching was chosen to be the bacilli or cocci classification method for chains and clusters in the final system. The main reason was because the method worked well for detection of cocci. Since only a distinction between cocci or bacilli was requested the method could be used as an exclusion method.

5.8.3 Classification of bacilli or cocci in the final system

The results of the classification of bacilli and cocci in the final system are presented in Table 4.8. The method in the final system combined measuring the major and minor axis of individual bacteria with the template matching function for chains or clusters. The system was better at detecting cocci than bacilli due to the reasons discussed above.

5.9 Future Challenges

The major challenge is the small size of the bacteria. Better resolution of the images would be of great importance when improving the system. However, the digital microscopy CellaVision uses today, has the highest resolution possible for a light microscope. There are methods for adding many images at different focus levels to get a sharper image. This could potentially help to make the images easier to work with and would not require any changes of the current system CellaVision uses today.

Bacteria are grouped into seven different types of bacteria, the most common types are Gram positive cocci, Gram negative cocci, Gram positive

bacilli and Gram negative bacilli which can be classified with this system. The other three less common types are corkscrew-shaped bacteria, bacteria with no cell wall and bacteria that only live inside other cells [39]. The method as it is today is limited to the first four types. A future bacteria detection and classification system should ideally support all seven groups.

5.10 Future Prospects

There are several benefits for patients and laboratory professionals that could be gained from an automatic bacteria detection and classification system. One is the benefit of digital images, another is the potential of increased speed a digital automatic method could provide.

Digitalization of the images provides a better and easier way to store the information from a sample rather than to keep the actual glass slide. If there is a need for a second opinion about the analysis, the images can easily be viewed later. If an analysis has been incorrectly made it is also easier to re-analyse the samples in retrospect. Moreover, a problem with keeping the slides instead of digitally storing images is that the samples can fade as mentioned earlier in the report (see section 4.1).

The digital storage of microbiology images also opens up for the possibility of a large data bank. This data bank could be used by researches and could potentially result in a greater understanding of diseases caused by microorganisms and bacteria resistance.

Digitalization of microscopic images can also improve the ergonomics for the laboratory staff. Working with a microscope oftentimes results in poor working positions. Since the bacteria would appear on a screen the problem with bad ergonomics could be reduced. This could help the overall health of the laboratory staff and reduce health costs due to poor working positions.

An automatic system can increase the speed of detecting potential bacteria when a bacterial infection is suspected. This could potentially reduce the unnecessary use of antibiotics. If the type of bacteria is determined faster and with more precision, a more specific antibiotic treatment could be given.

5.11 Conclusions

An overall method for the different steps: segmenting, detecting, and classifying Gram stained bacteria, as stated in the aim, has been developed. Different methods have been evaluated and tested for each step. The results from the method proposed implies that an automatic or semi-automatic system for detecting and classifying bacteria is possible to implement.

Chapter 6

Ethical Aspects

In medicine and technology many aspects of ethics can be discussed. One aspect is to protect the personal integrity, in particular when patient data is used in research. The data used in this project was anonymized by the personnel at the Microbiology laboratory in Lund and the identities of the patients were never revealed. It is important to keep the data anonymous to make sure there is no risk of discrimination in treatment depending on for instance the patient's age, gender and social status. In addition, the integrity of the patients could be violated if the data was not anonymous.

Other ethical aspects of digital pathology is that geographical discrimination can be decreased. By scanning the samples the data can be sent anywhere and in that way a physician in a larger city where specialists are more common can diagnose someone far out on the country side, as long as the small clinic has the needed equipment. The diagnosis could even be made over country boarders and could potentially help developing countries where specialists might not be available.

Bibliography

[1] Rydberg J. personal communication. Feb. 2016.

[2] Albrecht T Almond JW Alfa M Alton GG Aly R Asher DM et al. *Medical Microbiology*. 4th ed. University of Texas Medical Branch at Galveston, 1996.

[3] Chayadevi ML Raju GT. "Data mining, Classification and Clustering with Morphological features of Microbes". In: *International Journal of Computer Applications* 52 (2012), pp. 1–5.

[4] Mohan SK. *Gram Stain: Looking Beyond Bacteria to Find Fungi in Gram Stained Smear*. AuthorHouse, 2009.

[5] McClane BA Mietzner TA Dowling JN Phillips BA. *Microbial Pathogenesis*. Fence Creek publishing, 1999.

[6] Marler LM Siders JA Allen SD. *Direct Smear Atlas: A monograph of Gram-stained Preparations of Clinical Specimens*. Lippincott Williams & Wilkins, 2001.

[7] Otto M. "Staphylococcus epidermidis – the "accidental" pathogen". In: *Nat. Rev. Microbiol.* 7.8 (2009), pp. 555–567.

[8] *Bad Bug Book, Foodborne Pathogenic Microorganisms and Natural Toxins*. Second Edition. Food and Drug Administration, 2012.

[9] Brown J Hammerschmidt S Orihuela C. *Streptococcus Pneumoniae: Molecular Mechanisms of Host-Pathogen Interactions*. Elsevier, 2015.

[10] Elliott T Casey A Lambert PA John J. *Lecture Notes: Medical Microbiology and Infection*. Wiley & Sons, 2012.

[11] Donnenberg M. *E. coli: Genomics, Evolution and Pathogenesis*. Academic Press, 2002.

[12] *Sjukdomsinformation om Haemophilus influenzae-infektion (invasiv).* Folkhälsomyndigheten. 2016. URL: https://www.folkhalsomyndigheten.se/amnesomraden/smittskydd-och-sjukdomar/smittsamma-sjukdomar/haemophilus-influenzae-invasiv/.

[13] Mitchell JL Hill SL. "Immune Response to Haemophilus parainfluenzae in Patients with Chronic Obstructive Lung Disease". In: *Clin Diagn Lab Immunol* 7.1 (2000), pp. 25–30.

[14] Macenko M Niethammer M Marron JS Borland D Woosley TJ Guan X et al. "A method for normalizing histology slides for quantative analysis". In: *IEEE International Symposium on Biomedical Imaging* (2009), pp. 1107–1110.

[15] Tomasi C. *Orthogonal Matrices and the Singular Value Decomposition.* Duke University. 2013.

[16] Spiegelman M. *Applications of the SVD.* Columbia University. 2002.

[17] Johnston DA Ruifrok AC. "Quantification of histochemical staining by color deconvolution". In: *Anal. Quant. Cytol. Histol.* 23 (2001), pp. 291–299.

[18] Poppe R. *Segmentation.* URL: http://www.cs.uu.nl/docs/vakken/ibv/reader/chapter10.pdf.

[19] Otsu N. "A Thereshold Selection Method from Gray-Level Histograms". In: *IEEE Transactions on systems, man, and cybernetics* 9.1 (1979).

[20] Harrabi R Braiek EB. "Color image segmentation using multi-level thresholding approach and data fusion techniques: application in the breast cancer images". In: *EURASIP Journal on Image and Video Processing* 11 (2012), pp. 1–11.

[21] Shreve M. *RGB to HSI.* http://www.cse.usf.edu/ mshreve/rgb-to-hsi.

[22] Mignotte M. "Segmentation by Fusion of Histogram-Based K-Means Clusters in Different Color Spaces". In: *IEEE Trans Image Process* 17 (2008), pp. 780–787.

[23] Dempster AP. "Upper and lower probabilities induced by a multivalued mapping". In: *The Annals of Mathematical Statistics* 38 (1967), pp. 325–339.

[24] Shafer G. *A Mathematical Theory of Evidence.* Princeton University Press, 1977.

[25] Chaabane B Sayadi M Fnaiech F Brassart E. “Colour image segmentation using homogeneity approach and data fusion techniques”. In: *EURASIP Journal on Image and Video Processing* (2010), pp. 1–11.

[26] Chaabane B Sayadi M Fnaiech F Brassart E Betin F. “A new method for the mass functions estimation in the Dempster-Shafer's evidence theory: application to color image segmentation”. In: *Circ Syst Signal Process* 30 (2011), pp. 55–71.

[27] Efford N. *Digital Image Processing: A Practical Introduction Using Java*. Addison Wesley, 2000.

[28] Burger W Burge MJ. *Principles of Digital Image Processing: Fundamental Techniques*. Springer-Verlag London, 2009.

[29] *Morphological dilation and erosion*. MathWorks. 2016. URL: `http://se.mathworks.com/help/images/morphological-dilation-and-erosion.html`.

[30] Haralick RM Shapiro LG. *Computer and Robot Vision*. Vol. 1. Addison-Wesley, 1992.

[31] Zhang G Ma Z. M Tong Q He Y Zhao T. “Shape Feature Extraction Using Fourier Descriptors with Brightness in Content-based Medical Image Retrieval”. In: *International Conference on Intelligent Information Hiding and Multimedia Signal Processing* (2008), pp. 71–74.

[32] *vision.TemplateMatcher System object*. MathWorks. 2016. URL: `http://se.mathworks.com/help/vision/ref/vision.templatematcher-class.html`.

[33] *Template Matching*. MathWorks. 2016. URL: `http://se.mathworks.com/help/vision/ref/templatematching.html`.

[34] de Berg M. *Computational Geometry: Algorithms and Applications*. Springer Science & Business Media, 2000.

[35] Kong TY Rosenfeld A. *Topological Algorithms for Digital Image Processing*. Elsevier, 1996.

[36] Wu Q Merchant F Castleman K. *Microscope Image Processing*. Academic Press, 2010.

[37] Meyer F. “Topographic distance and watershed lines”. In: *Signal Processing* 38 (1994), pp. 113–125.

[38] Vincent L Soillle P. “Watersheds in Digital Spaces: An Efficient Algortihm Based on Immersion Simulation”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 13.6 (1991).

[39] Senior K. *How Many Types of Bacteria Are There?* Feb. 2016. URL: http://www.typesofbacteria.co.uk/how-many-types-bacteria-are-there.html.

Master's Theses in Mathematical Sciences 2016:E23
ISSN 1404-6342

LUTFMA-3296-2016

Mathematics
Centre for Mathematical Sciences
Lund University
Box 118, SE-221 00 Lund, Sweden

http://www.maths.lth.se/