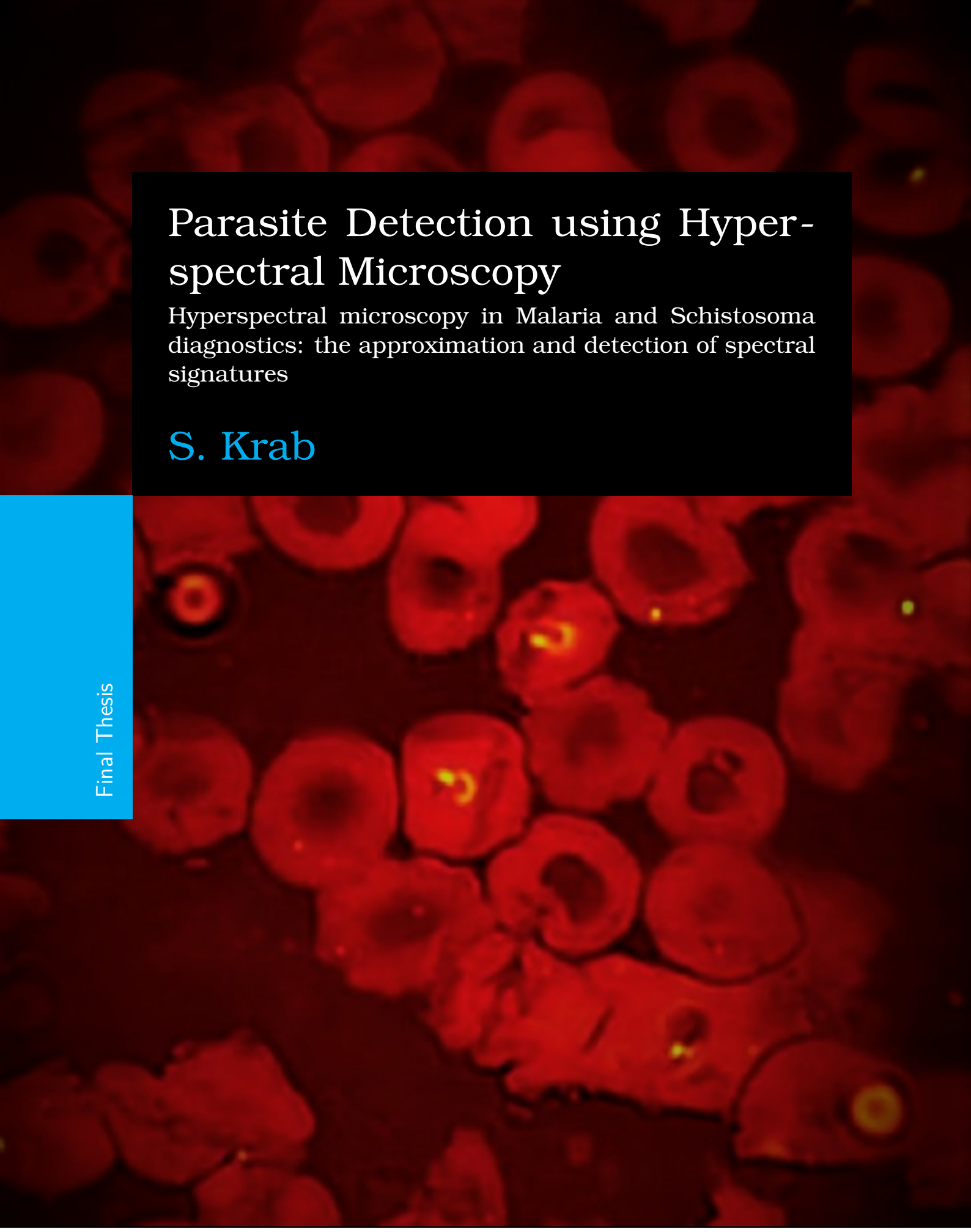


Parasite Detection using Hyperspectral Microscopy

Hyperspectral microscopy in Malaria and Schistosoma diagnostics: the approximation and detection of spectral signatures

S. Krab

Final Thesis

Delft Center for Systems and Control

Parasite Detection using Hyperspectral Microscopy

Hyperspectral microscopy in Malaria and Schistosoma diagnostics: the approximation and detection of spectral signatures

Final Thesis

S. Krab

November 4, 2021

Faculty of Mechanical, Maritime and Materials Engineering (3mE) · Delft University of Technology

Abstract

Parasitic diseases such as malaria remain a mayor burden on global health. One of the biggest challenges still to be overcome is that of inadequate diagnoses. This research explores the opportunities that Hyperspectral Imaging yields in this field. The first goal is to estimate the spectral signature of Malaria parasites in non-stained or Giemsa-stained thin smear blood samples and of Schistosoma parasite eggs in urine samples. For this different endmember extraction algorithms are combined with various methods of pre-processing and dimensionallity reduction. The used endmember extraction methods are pure pixel index (PPI), NFINDR, Statistics Based and simplex identification via split augmented Lagrangian (SISAL). For denoising Savitzky Golay and 3 dimensional gaussian filtering is used and the dimensionallity reduction is done with PCA, ICA or HySime. The resulting spectral signatures of the algorithms are validated by inspecting the endmember locations, spectra and abundance maps. They have furthermore been compared by the classification performance where the spectral signatures are used in the feature derivation. This is done by deriving a detection map using OSP or CEM detection and then using the SVM or random forest classifiers to classify cells as being infected or not. These performances are furthermore compared to RGB image based classification.

In case of the stained Malaria sample the four endmember extraction methods are shown to be applicable to various degrees. Firstly, the PPI method is shown to be inconsistent, resulting in different spectra each run. Secondly, the statistics based method unable to separate the spectral signatures of the red blood cells and thirdly the background. Thirdly, The NFINDR method seems to work well considering the endmember locations, spectra and abundance maps, but leads to a low classification performance. The research concludes that Sisal made the most accurate estimations of the spectral signature of the parasite. The results from all the validation methods are in line with expectations. Furthermore, the use of this spectral signature in the feature derivation process results in the highest classification performance. This performance is also shown to be significantly higher compared to using either the first principal component of the full hyperspectral data or the RGB images. Applying the same methods to the Schistosoma sample it is found that some of the methods, though interestingly not Sisal, are able to to create an abundance map in which the egg is separated from the background. However, none of them are able to separate the egg and the white blood cell

and detection maps using these signatures did not show the egg more clearly than the first principal component did. None of the methods are found to be able to extract the spectral signature of the unstained Malaria parasite.

Finally, a hypothetical multispectral microscope is proposed which images at the wavelengths where the spectral signature of the parasite in a stained sample has the biggest difference in light transmittance to the other endmembers. This setup is simulated from the available hyperspectral data and its classification performance is compared to classification performance using the full hyperspectral data and using the RGB images which are simulated from the same data. The classification using the discriminative wavelengths is found to outperform both in terms of sensitivity and specificity. This implies that the images at these specific wavelengths provide more discriminative power and such a multispectral setup could provide a significant advantage over RGB imaging.

This master thesis is performed at Delft Center for Systems and Control (DCSC), Delft University of Technology (TU Delft).

Table of Contents

List of Figures

Preface

This document is my Master of Science graduation thesis. I have chosen to do my research on the subject of parasite detection using hyperspectral imaging. The subject specifically appealed to me since it combined my interest in optics and data science with a prospect of contributing to a field of research that aims to prevent the deaths and sorrow of millions of people. This has given me a clear purpose and motivator and my hope is to provide a significant contribution to the advance of parasite detection techniques with my research.

Acknowledgements

I would like to thank Gleb Vdoving for providing me with a chance to do my master thesis in an incredibly interesting field and for his assistance during the writing of this report. I'd furthermore like to thanks Tope Agbana for his assistance and for his idea to test the algorithm on schisto egg samples as well as the malaria samples, which gave another dimension to my research. I would furthermore like to thank Casper van Engelenburg for helping me get started with both the hardware and software side of the microscope setup.

Delft, University of Technology S. Krab
November 4, 2021

"Man is nothing else but what he makes of himself."
— *Jean-Paul Sartre*

Chapter 1

Introduction

Parasitic diseases such as Malaria and Schistosomiasis remain a mayor burden on global health. The majority of the deaths related to these diseases occur in sub-saharan Afrika where there is a lack of good health facilities. This often leads to inadequate diagnoses which results in the disease progressing further than necessary. The lack of proper training of the microscopists has been proven to be one of the root causes. As an answer to this problem the field of autonomous detection arose, aiming to automate this part of the diagnostics process. This has traditionally often been done by using bright field microscopy, capturing the image using and RGB sensor. After segmenting the cells from the image some classifier is trained to determine whether a cell is infected or not. Most research these days is directed at the application of different segmentation and classification methods. This research aims to explore a different path to improve parasite detection, namely multi- and hyperspectral imaging. These imaging methods have found popularity in a wide range of applications over the last decade due to the ever increasing computational power at our disposal. The increased spectral resolution in hyperspectral images gives a lot more data to work with. This data could be used directly in the process of classifying cells to be infected or not. Given a big enough samplesize this is likely to result in improved classification performance. However, as the samplesize in this study is rather small, using the data directly will likely lead to overtraining. The direct application of hyperspectral data in classification will therefore not be the main aim of the study. Instead, the main aim will be to use the hyperspectral data to estimate the spectral signatures of the parasites using various endmember extraction methods. A somewhat large amount of combinations, namely PPI, NFINDR, statistics based and Sisal combined with multiple different denoising and dimensionallity reduction methods, will be tested. This is because little is known about the application of endmember extraction in microscopy. As the methods have different strengths and weaknesses each of the different types of endmember extraction is explored. This leads to the first research question

Can Hyperspectral Imaging be used to accurately determine the spectral signatures of the Malaria and Schistosoma parasite?

However, hyperspectral imaging equipment is rather expensive and thus not very suitable for direct application in sub-saharan Afrika. Nonetheless, it might prove valuable in that the spectral signatures of the parasite and the other substances in the sample can be used to determine the wavelengths which have the most discriminative power. This could in help with the development of a more affordable multispectral machine with greater capabilities than its RGB based counterpart. the second research question will thus be

Does Multispectral Imaging based on these spectral signatures provide a significant benefit over traditional RGB imaging in autonomous detection?

1-1 Outline

The report is made up of 4 more chapters. In the first chapter Malaria and Schistosomiasis are introduced along with some of its relevant features, followed by a general overview of some of the more popular methods of red blood cell segmentation and classification as of the moment of writing. Here, the methods are explained in the context of RGB images, but note that these methods can be applied to any image. The main aim of this chapter is to give a good introduciton into autonomous parasite detection. The next chapter gives an introduction into hyperspectral imaging, after which the all the steps regarding endmember extraction are treated in depth. Several algorithms on virtual dimensionallity estimation, dimensionallity reduction and endmember extraction are discussed. After this chapter come the results. First, the classification using the hyperspectral data versus the RGB data are compared. Next, the performance of the endmember extraction methods are validated by inspecting the corresponding pixels, spectra and abundance maps. The methods are furthermore compared by their classification performance using the signatures to create detection maps. In the last experiment, a hypothetical multispectral setup is tested against a RGB setup to determine whether the previously mentioned discriminative wavelengths do provide an advantage over RGB imaging. In the final chapter the results are discussed and a conclusion is given in relation to the research questions.

Chapter 2

Autonomous Parasite Detection

This chapter will give an short introduction into the current state of Malaria and to a lesser extend Schistosomiasis. It will cover the parasites that causes the diseases and the most widely used methods of diagnosis. It will go on to discuss the method of conventional bright field microscopy as it remains the gold standard and is the most closely related to hyperspectral imaging. The main body of this chapter will subsequently focus on autonomous detection using bright field microscopy, focusing mostly on Malaria as this research field is much larger. It will provide the reader with insight into the current state of research and its shortcomings. The subjects of preprocessing, red blood cell segmentation and finally classification are touched upon and should give the reader necessary background knowledge and serves as a good introduction into the main subject of the research. Reading this chapter it will become clear to see the potential hyperspectral microscopy has in this field.

2-1 The State of Malaria Worldwide

Malaria is a parasitic disease that according to the World Malaria report, though declining, still costs the lives of an estimated 405000 people worldwide in 2018. 2013 million of the total 228 million cases happened in Africa [16]. Most of these deaths occur in rural areas and most researchers agree that a lack of good health facilities plays a big role. Inadequate diagnosis is one of the mayor problems that still has to be overcome.

2-1-1 Malaria Detection

Malaria can be diagnosed accurately in more developed countries. Here, bright field microscopy using giemsa stains remains the gold standard. It has high sensitivity, the chance of a positive sample to test positive, high specificity, the chance of a negative sample to test negative, and a low limit of detection, meaning the threshold for amount of parasites in the blood for them to get detected, is small. However, this method is very time consuming and requires trained personnel, lack of which can lead to vast variations in performance [17][18].

In response to this problem various rapid diagnostic tests (RDT) have been developed. These are fast and easy to perform. However, as described in [19], though outperforming poorly executed bright field microscopy diagnoses, it has its limitations. Sensitivity has been reported to vary widely [20] and the tests are also unable to detect mixed infections, distinguish between species and detect low concentrations of parasites. The various other methods that have been developed also tend to fall into one of these these two categories, either requiring a lot of expertise and being time consuming or having severe limitations. In recent years, this has led to the development of various methods combining microscopy and image analysis software and/or machine learning algorithms [2], ranging from conventional classifiers to neural networks. These methods can greatly improve the reliability of diagnoses as they are not reliant on the skill of the microscopist. They can furthermore speed up the process allowing for more patients to be tested and could reduce cost.

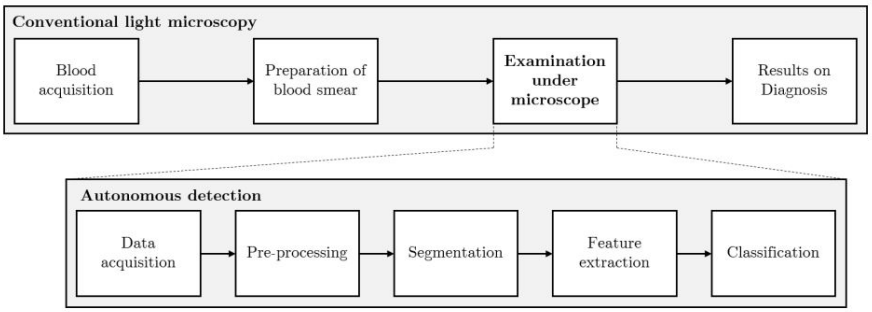


Figure 2-1: Illustration of how the manual examination under the microscope is replaced by an autonomous process of data acquisition, pre-processing, cell segmentation, feature extraction and finally classification [1]

2-1-2 The Plasmodium parasite

Malaria is a disease caused by parasites that belong to the Plasmodium genus. In humans, Malaria is caused by five different Plasmodium species named P.falciparum, P.vivax, P.Malariae, P.ovale and P.knowlesi. Of these P.falciparum and P.vivax account for the mayority of deaths. The different types will not be treated individually, but one should be aware that being able to differentiate between them is beneficial in determining the best treatment. These parasites are transmitted through bites of infected female Anopheles mosquito's. In humans, the parasites first grow in the liver and then spread to the red blood cells where they start to cause symptoms. In the blood the parasite infects a red blood cell where it undergoes four stages, each giving a distinctly different appearance under a microscope. The total development cycle takes about 48 hours during which the parasite goes through the ring, trophozite, schizont and gametocyte stage. After the gametocyte stage the cell is destroyed and merozoites are released which then go on to invade other cells starting over the cycle [21]. Being able to differentiate between stages can be advantageous as it is a good indicator of the severity of the disease. Non-severe cases mostly showing first stage parasites and severe cases often showing all stages to be present. Figure 2-2 gives microscope images of each of the species and stages in a stained thin smear sample as well as some characteristics to differentiate between types.

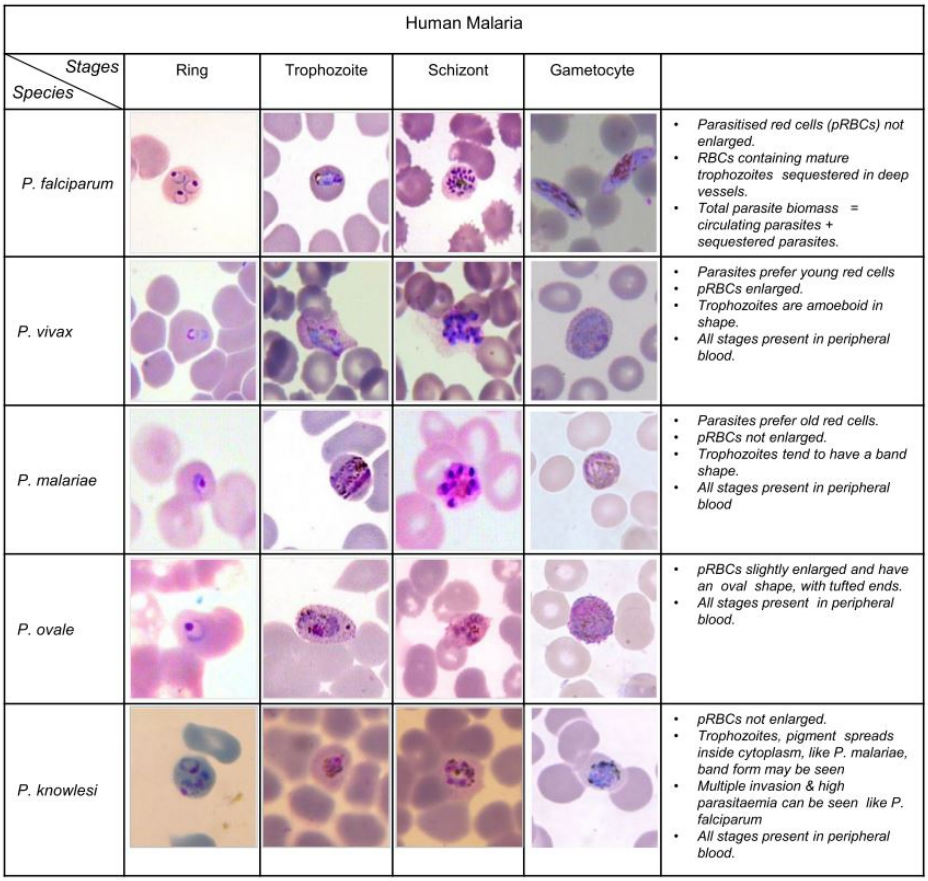


Figure 2-2: Five different human Malaria Plasmodium species and their life stages in thin blood film [2] (Source: K. Silamut and CDC)

2-2 The State of Schistosomiasis Worldwide

Schistosomiasis is a parasitic disease that is caused by trematode worms. The parasite spreads through infested waters and it is estimated that some 236.6 people are at risk of the disease and the estimated death toll varies between 24072 and 200000. Almost half of the people at risk receive preventive treatment as of 2019. As was the case with Malaria, most cases occur in sub-saharan Afrika. Similarly, inadequate diagnosis remains one of the biggest problems.

2-2-1 Schistosomiasis Detection

For Schistosomiasis there are two main approaches to diagnostics, antibody tests and bright field microscopy, both with some mayor drawbacks. The antibody test is very fast and sensitive, but the antibodies remain present for months to years after infection, the test thus not being able to distinguish between current and past infections. In case of bright field microscopy, the disease cannot be diagnosed for the first two weeks, after which the eggs start to show up in the faecal matter. Still, this has resulted in the microscopy based method being used in most areas. As is the case with Malaria, using bright field microscopy, Schistosomiasis

can be accurately diagnosed given careful procedure and trained personnel. In this case faecal matter, using the Kato-Katz technique, or urine samples are examined under the microscope. If the person is infected, both will contain the eggs of the parasite. In case of the Kato-Katz technique a piece of cellophane which has been soaked in methylene blue glycerol is used to make the eggs visible, while in case of the urine test the eggs are generally made visible by adding iodine. It is however still necessary to use a 20x or 40x lens to see the eggs, which makes the process of examining the whole sample rather slow. Using lower magnification lenses would make it possible to examine larger areas of the sample at the same time, but it also makes it significantly harder to spot the eggs. Some papers show classifiers using textural information [22]. Others show convolutional neural networks [23] can be applied to this task to reasonable success, but much is still to be won.

2-2-2 The Schistosomiasis Parasite

The Schistosomiasis disease is caused by blood flukes, parasitic worms, of the Schistosoma genus. There are 2 major forms of Schistosomiasis, intestinal and urogenital. Of the intestinal kind there are four four species, the Schistosoma mansoni, Schistosoma japonicum, Schistosoma mekongi and Schistosoma guineensis. Of the urogenital kind there is one, Schistosoma heamatobium. Each specie has a distinct form and size, as displayed in figure 2-3. In this research we'll be focusing on the intestinal species as these result in eggs in the urine, thus being detectable in the way described before. Infection occurs when larval forms of the parasite, released by freshwater snails, penetrate the skin while in contact with infested water. It can furthermore be transmitted among humans infected humans contaminate water sources with their excreta containing the parasite eggs. When the parasite has entered the body it further develops. The females then lay eggs which end up in faeces and urine, as well as body tissue, where it causes harm to the organs of the host.

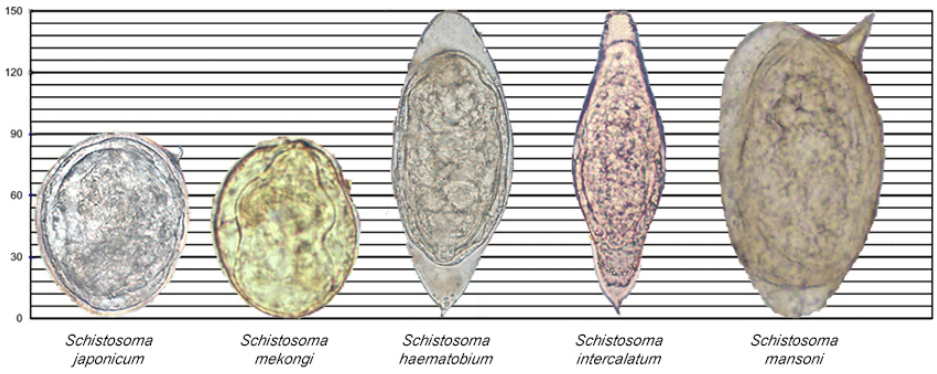


Figure 2-3: The different species of Schistosoma, size given in micrometers

2-3 Light and Microscopy

Understanding the diseases, the next step is to understand the way they are detected. The most commonly used method for Malaria detection and one of the more common once for

Schistosoma detection is bright field microscopy and it remains the gold standard to this day. The theory of how light interacts with matter and how this is used in bright field microscopy will be treated somewhat more in depth in this section as this is knowledge is important for understanding hyperspectral microscopy later on.

2-3-1 Light Absorption, Reflection and Transmission

Crucial to understanding microscopy is understanding why we see what we seen when we look into the microscope. As light hits matter, in this case the sample, one of several things can happen. The light is either reflected, absorbed, scattered or transmitted, each illustrated in 2-4. In case of reflection the light bounces of the new medium under the same angle. Light absorption happens when the new medium takes up the energy of the photon. Scattering is somewhat related as the light is first absorbed and then re-emitted, often in a different direction and wavelength. Finally, when light is transmitted it passes through the new medium (albeit its direction is changed, see 2-3-3). Of course, not one, but a combination of all of these phenomena happen simultaneously, and in different proportions depending on the wavelength of the light. In a microscope the sample is homogeneously lit from below and only the transmitted light is observed. Due to the difference in light transmittance of substances at different wavelengths it is possible to differentiate between them.

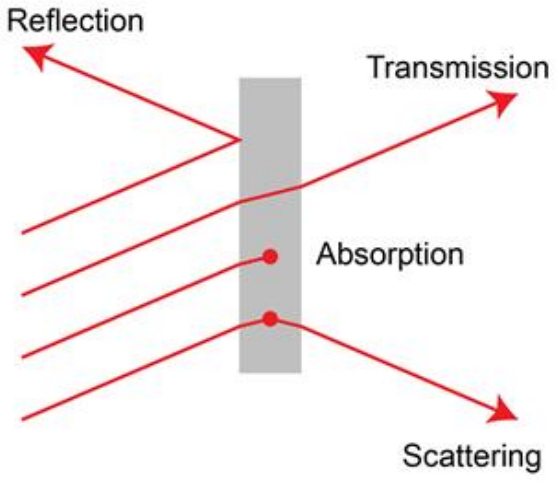


Figure 2-4: The four main interactions light has with matter

2-3-2 Light as a Wave

Without getting to in depth, as this is not necessary to understand the rest of the research, light in optics is treated as wave in the electromagnetic spectrum, the electric and magnetic field of this wave being orthogonal to each other and propagating away from a point source. Its respective wavelength is determined by the source that produced it and different wavelengths have vastly different effects on matter. These wavelengths range from 10000km - 1mm (radio waves) to smaller than an atomic nuclei (gamma rays), with the visible spectrum in between ranging from 400 to 700 nm. Treating light as a wave in the electromagnetic spectrum helps

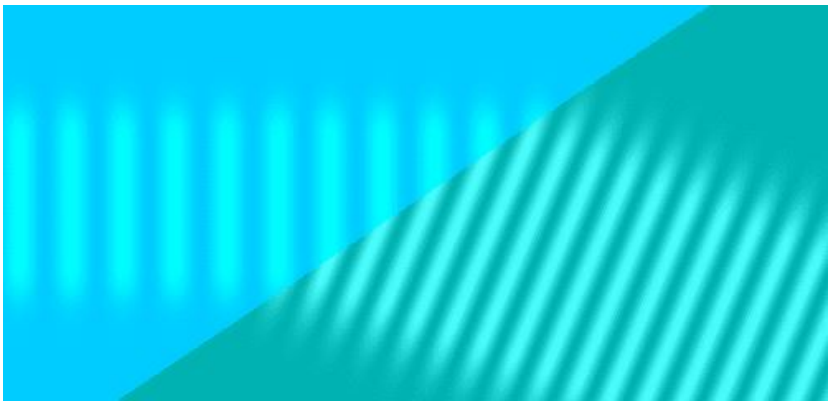


Figure 2-5: Refraction of light at when changing medium [3]

us understand some essential phenomena in microscopy, namely refraction and the Rayleigh criterion.

2-3-3 Refraction

Refraction can be understood by looking at what happen to the electromagnetic wave at the plane of incidence as it enters a new material. When entering a new material, the speed of propagation of the wave changes, which in turn causes the the waves in the new material to have a different wavelength. However, since the waves need to be connected at the plane of incidence, the direction of the wave must change (unless the wave is perpendicular to the plane of incidence), as can be seen in 2-5.

The refraction change of direction due to this refraction can be most easily calculated using Schnell's law, given by

$$\frac{\sin\theta_1}{\sin\theta_2} = \frac{n_1}{n_2}$$

where θ_1 is the angle between the direction of propagation of the light in the first medium and the plane of incidence and θ_2 the angle between the direction of propagation of the light in the second medium and the plane of incidence. n_1 and n_2 are the refractive index of medium 1 and medium 2.

2-3-4 Lenses

Lenses make use of this phenomenon. Due to their shape the change in direction of propagation of the light happens twice in the same direction, causing a coherent light source to either be converging or diverging afterwards. In case of a converging lens, the coherent light source is focused in a single point at the focus length of the lens. Any point source is also focused at a certain distance from the lens which can be calculated using the thin lens formula.

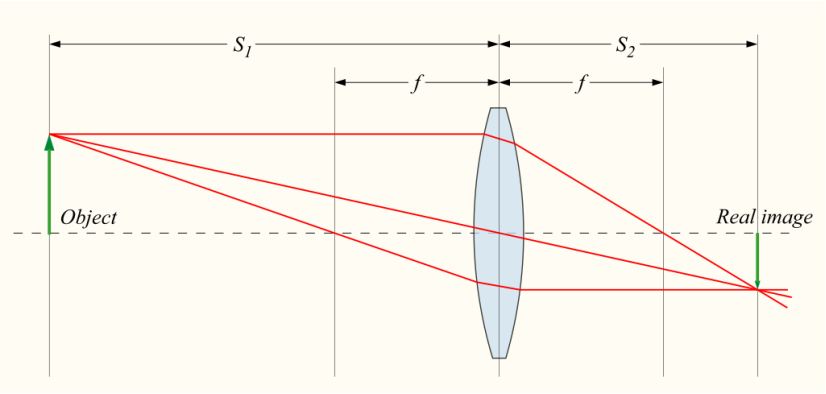


Figure 2-6: A real image of the object is created at a distance which can be determined using the thin lens formula [4]

$$\frac{1}{f} = \frac{1}{s_1} + \frac{1}{s_2}$$

Here, f is the focal length of the lens, s_1 is the distance of the point source to the lens and s_2 is the distance to the plane in which the image of the point source is formed. As an object can for the purpose of optics be represented as a multitude of point sources, a real, inverted image of the object will be formed in this plane, as is represented in **??**.

2-3-5 Rayleigh Criterion

However, these light waves interact with each other when passing through a circular aperture or lens. Two waves of the same amplitude, wavelength and polarization with no phase difference would add up in what is known as constructive interference, while two waves of the same amplitude, wavelength, polarization, but a 180 degrees phase difference would cancel out completely in what is known as destructive interference. Due to this interference the light source is not reconstructed perfectly, but produces a ring shaped diffraction pattern, known as the airy disk. The radius of the inner circle can be calculated using

$$\theta = 1.22\frac{f\lambda}{D}$$

where is the wavelength, f is the focal length of the lens and D is the diameter of the lens. When two points are too close together, their airy disks will be indistinguishable from each other, thus limiting the spatial resolution of the image, as can be seen in 2-7.

2-3-6 Microscopy

A microscope uses multiple optical elements such as lenses and apertures to first properly illuminate the sample and then enlarge it so it can either be looked at directly or captured by a

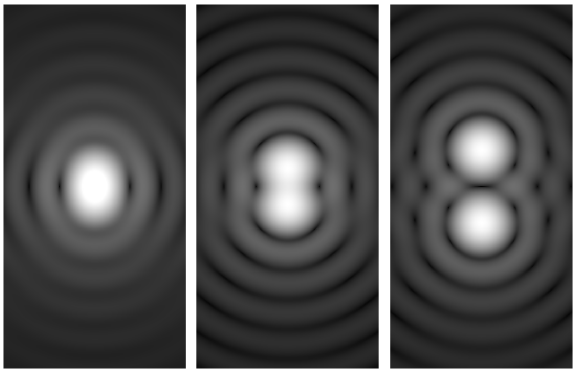


Figure 2-7: Two airy discs located too close together are indistinguishable from each other

camera sensor. The schematic **??** shows the microscope used in this research. The microscope is broken up into three sections, the illumination path (orange), the sample (green) and the imaging section. Given all the optical elements are aligned properly the light passes through through the various lenses in the illumination path eventually homogeneously illuminating the area of interest on the sample with coherent light. Moving the objective lens such that the sample is in the focal plane it then creates a coherent beam which can either go through an eyepiece which makes it possible to observe the sample directly or be focused on the camera sensor by the tube lens as is the case in this setup.

2-4 Sample Preparation

Focusing back on Malaria, before the sample can be observed under the microscope it needs to be prepared. There are multiple choices to be made here. The first important choice is between creating a thin or thick smear sample, each with their own advantages and disadvantages. The second choice what type of staining is to be applied to the sample to make the parasites more visible. This is then furthermore dependant on the type of microscopy used. This section will give a short overview of some of the possible choices.

2-4-1 Thin and Thick smear samples

In order to observe the blood sample under the microscope a drop of it needs to be applied to a glass slide. This can be done in one of two ways, a thick or thin smear, the effects of which can be seen in 2-8. Using the thick smear it is currently only possible to detect whether the parasite is present, while using the thin smear it is also possible detect the type and stage of the parasites, as well as the parasitemia, the percentage of cells that are infected. These factors play an important role in the determination of the severity of the case and in choosing the right treatment. However, thick smears are generally used in the field as they make it possible to examine a larger amounts of blood cells at once, increasing the sensitivity[24].

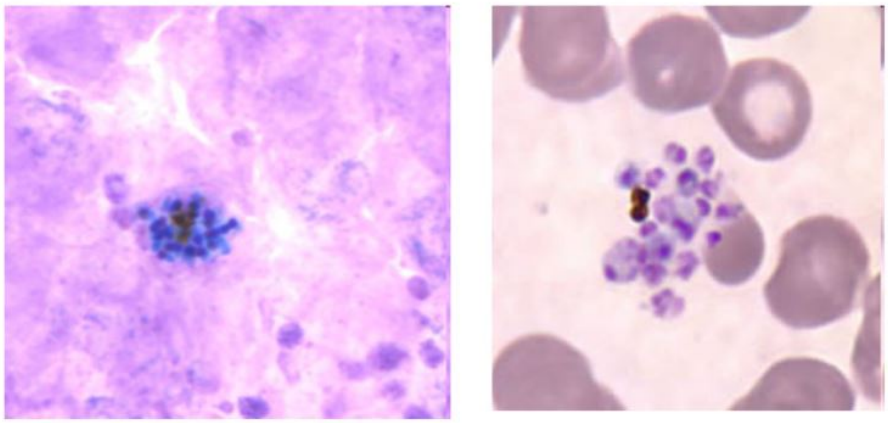


Figure 2-8: A thick (left) and thin (right) smear blood sample under the microscope [5]

2-4-2 Staining

The next topic to discuss is staining. Hyperspectral imaging could possibly make this step redundant, but for traditional methods it is necessary in order to be able to see the parasites. Many different kinds of stains have been developed over the years in order to increase detection performance. However, the most widely used stains remain the Giesma (1902) and Leishman (1901) stain combined with bright field microscopy. The former has been proven to be very reliable but requires experienced personnel and is relatively time consuming while the latter is cheaper and easier to perform at the cost of being slightly less reliable. Furthermore, several staining methods have been developed for other types of microscopy and have been found to be advantageous in some regards. Notable examples are fluorchrome [25] and Dapi/Mitotracker [26]. using the fluorchrome stain for example, the sample is illuminated by a near monochromatic (single wavelength) coherent light source with a very specific short wavelength which is absorbed by the fluorophores (fluorescent stains) which then emit light of a much longer wavelength in the near infrared range. The source light can then be filtered out to reveal the materials which the fluorophores bind to.

2-5 Data Acquisition

Traditionally the Malaria detection is mostly done by direct observation. The sample is observed under the microscope to make a diagnosis. However, the last few years have seen a lot of research is being done to automate this last step, capturing the image on a camera and then automating the diagnosis. Depending on the microscope and method to be used, different kinds of cameras can be used. Bright field microscopy is the most used as the equipment necessary is cheap and therefore usable in low-resource settings. On the other hand fluorescent microscopy has shown to be very reliable in cases where type, stages and parasitemia needs to be determined [25]. These methods require different kinds of microscopes and imaging equipment. In bright field microscopy only the part of the electromagnetic spectrum that belongs to visible light is employed to image the sample. The sample is illuminated by coherent white light. An image is created by a RGB (red, green and blue) CCD camera

which measures an three intensities per pixel, one for each of the colorbands. These values can be used to recreate the color image of the object. For other kinds of imaging such as fluorchrome the sample is generally imaged by a monochromatic CCD camera producing a single intensity image as only the light at a specific wavelength is of interest.

2-6 Preprocessing

When the image has been captured by the camera sensor it generally first needs to be preprocessed. The main objectives here are to remove unwanted distortions such as noise and enhance features such as contrast and signal-to-noise ratio. Various relevant methods of spatial preprocessing are discussed in this section. Since they are of spatial nature these are generally applied to the intensity images of the three colorbands if bright field microscopy is used or the single greyscale images resulting from various other types of microscopy. One of the most commonly used spatial processing techniques is smoothing. Smoothing is often used to reduce noise. It does so by convolving the image with a mean or gaussian filter where the output pixel is the average or weighted average of the original pixel and neighbouring pixels. To increase contrast, two techniques are most commonly used. Contrast enhancement and (adaptive) histogram equalisation, the effects of which can be seen in 2-9. In the case of contrast stretching the range of intensity values of the image are stretched to the maximum dynamic range. In case of histogram equalization a linear mapping is found that brings the histogram in which intensity value is equally probable of occurring.

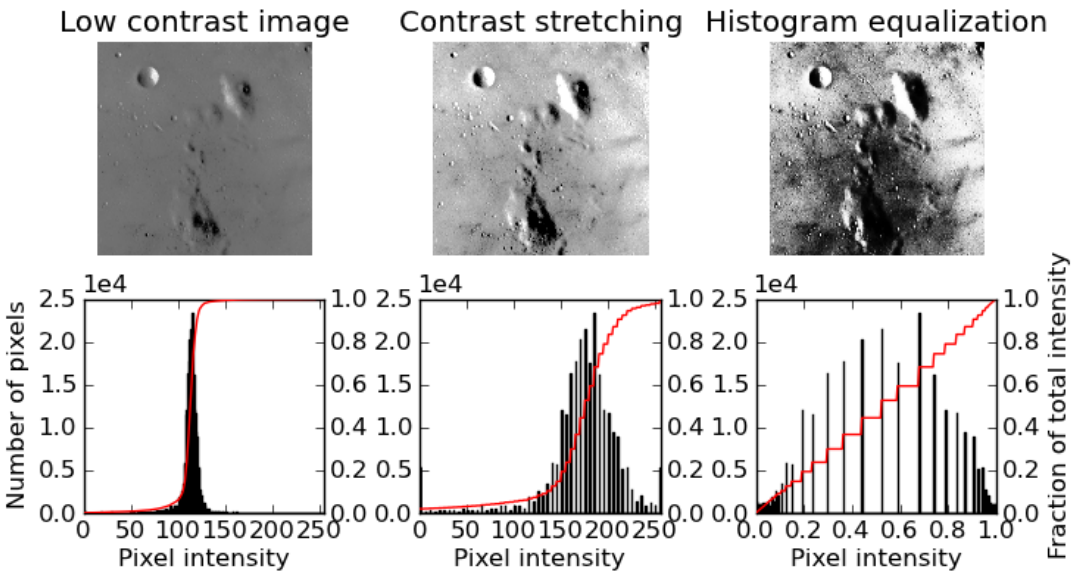


Figure 2-9: Effects of contrast stretching (middle) and histogram equalization (right) [6]

2-7 Red Blood Cell Detection and Segmentation

When the image has been enhanced it can be analysed. Generally the first step is red blood cell segmentation. This is the process that creates masks which are used to separate each cell

from the rest of the image, as shown in 2-10. In most publications some type of thresholding is used to create a binary image after which morphological operation are used to create masks that can be used to segment an area (in this case the blood cell) from the image. A popular type of thresholding is Otsu thresholding [27], which separates the image into a foreground and background by finding a threshold which minimizes intra class variance. Further processing can be done to split clusters, or these can simply be left out. However, in some publications edge detection is used instead. An edge of an object in the image is often a high fluctuation in intensity in one or several wavebands. However, since using these filters corresponds to high pass filtering, it also results in amplifying the noise in the image. These techniques may thus not work well in high noise datasets. Common methods in this category include gradient and Laplacian filters. These edge maps can be used to find boundary edges by some criterion, as in [28]. The terminal points of these boundary edges are then linked together if they are close to each other and the curvature is similar to the curvature of RBC's. As before, these closed circles can be further processed to split clusters and then be used to create masks. Both techniques are generally combined with morphological operation such as opening and closing in order to determine the final mask. If performed correctly, this will result in non connected masks that are close to the outlines of the RBC's. Each RBC can now be separated and classified.

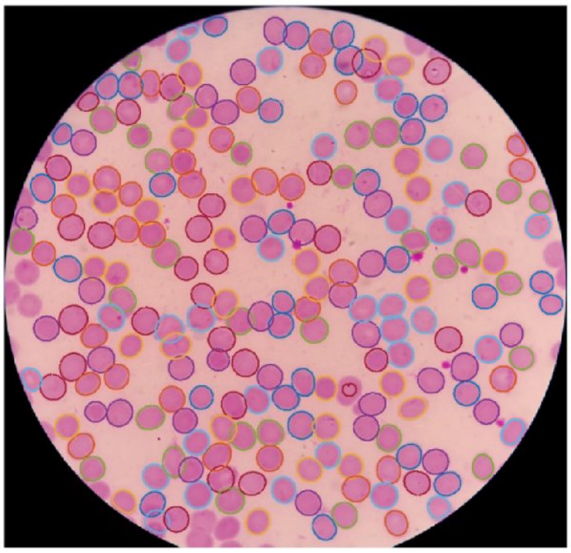


Figure 2-10: segmentation boundaries obtained in RBC segmentation [7]

2-8 Parasite Detection

Classifying the RBC's as being infected or uninfected requires the choice and training of a classifier. The training in turn requires a training set consisting of a preferably large set of examples for which the classification of being infected or not is known. such a set is generally created by a trained microscopist. However, in case of many traditional classifiers a small set of features needs to be extracted from the image of the RBC. These features should hold as much useful information on the cell as possible as they are the input by which the classifier

has to determine whether a RBC is infected or not. However, the choice of classifier also has a great influence on the choice of features.

2-8-1 Feature Extraction

Preferably, a set of features is determined which is not too large as this could lead to worse performance due to the curse of dimensionallity, a phrase introduced by [29] referring to various phenomena that occur in the analysis of high dimensional spaces. However, the set has to be chosen big enough and in such a way that it provides enough discriminative power to determine which RBC is infected and which is not. These features are generally derived from greyscale or colorband intensity images and are generally statistical, geometrical or textural in nature. In the first case a feature is to derived from the intensity histogram of an image containing the intensity values and their number of occurrences. These histograms can then be described by a few descriptors. Obvious possibillities are mean, variance and higher order statistics. Textural features, introduced by Haralick [30], try to capture the textural information in the image. It assumes that this textural information is contained in the overall average spatial relation between grey tones in the image. A grey level co-occurrence matrix is determined which counts the occurrences of neighbouring pixels to the reference pixels given a certain distance. This matrix is then used to determine the various Haralick descriptors. The most commonly used are maximum probability, correlation, contrast, energy, homogeneity and entropy as the other descriptors are all correlated to these. Finally, some geometrical features can be derived from the masking image. Common examples are area, perimeter, compactness and circularity.

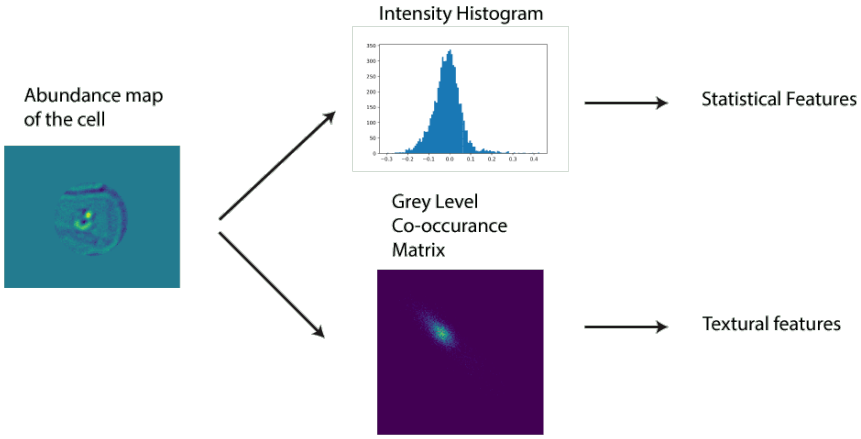


Figure 2-11: feature extraction diagram

2-8-2 Classification

Finally, having derived a feature set for each cell we can move on to the last stage where the infected and non-infected red blood cells are to be distinguished, as shown in 2-12. This is generally achieved with some sort of classifier. Examples of the use of supervised learning algorithms for Malaria detection such as the linear discriminant analysis or quadratic

discriminant analysis classifier, k-nearest neighbor classifier and support vector machine are easily found in the literature and often provide good results. The first two are based upon the assumption that the classes have a Gaussian density distribution. The k-nearest neighbour classifier does not make this assumption but does rely on euclidean distances for its classification. These classifiers are thus more likely to suffer from the curse of dimensionality. Less susceptible to this are support vector machines. These types of classifier have been getting more and more popular over the past few years due to the fact they can handle larger amount of features relative to the size of the dataset. The same benefits and often even better performance can be achieved with random forest classifiers, but these are computationally heavier. Lately, the field has also been expanding into neural networks and deep learning classifiers [31]. Though these have been shown to outperform traditional classifiers, they are very computationally heavy to train and need large datasets. As no publicly available dataset of hyperspectral Malaria sample images exists and needs to be created for this research, these types of classifiers will are not considered for this research. The classifiers used in this research are discussed more in detail at the end of the next chapter.

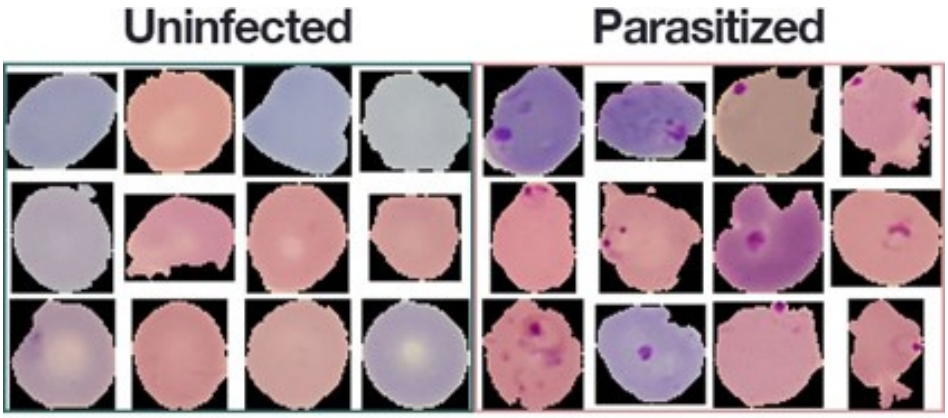


Figure 2-12: infected cells (left) and parasitized cells (right) according to classification algorithms [8]

Chapter 3

Hyperspectral Image Analysis

This chapter will start with a short introduction into how hyperspectral imaging (HI) works. It then proceeds to discuss how the preprocessing should be approached differently compared to RGB, after which it moves on to the main subject of the chapter, endmember extraction. In this process a certain amount of signal sources, which most likely make up the data, are estimated. This process consists of several stages. In the first stage the amount of distinct spectral signatures present in the data, namely the virtual dimensionallity (VD), is estimated. The second stage reduces the dimensionallity of the data. In the last step the endmembers are extracted from the data. The algorithms discussed in this chapter have each found themselves to useful in a range of applications. The chapter then moves on to how these endmembers can then be detected in new data and be used to make a per cell classification of whether it is infected or not. It finally concludes with a short overview of the microscope used in the experiments.

3-1 Introdcuction to Hyperspectral imaging

HI is a very powerful technique that has found popularity in a wide range of applications over the last decade. This rise in popularity is mainly caused by the decrease in price of the necessary equipment and ever increasing power of computers. These factors are making the acquisition and processing of the large amount of data in hyperspectral imaging much more achievable. HI differs from conventional color imaging in that instead an intensity value in three large wavebands, namely belonging to the red, green and blue (RGB) color ranges in the visible spectrum, it captures many more intensity values as it divides the electromagnetic spectrum into many more small wavebands. Due to this higher so called spectral resolution, it contains a lot more information.

3-1-1 Construction of the Hypercube

In HI, for each pixel an intensity value for many wavebands is captured. This results in a 3D hypercube of data, two spatial dimensions and one spectral dimension, as illustrated in 3-2.

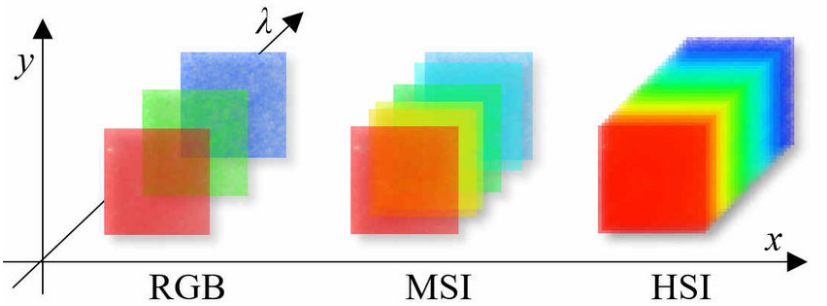


Figure 3-1: The difference between RGB, multispectral and Hyperspectral imaging

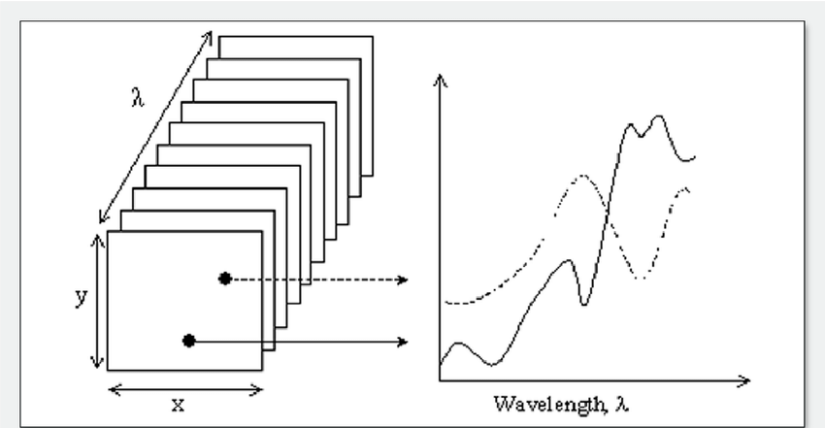


Figure 3-2: A schematic representation of the hyperspectral imaging hypercube showing spatial dimensions x and y and spectral dimension λ. Spectra on the right correspond to single pixel locations in the image [9]

An image containing the intensity values of a single waveband is called an intensity image and a vector containing the intensity values of all wavebands in a single pixel is called a spectrum. A hypercube can be constructed in 2 ways. The first method is to use a spectral sensor at each location to measure its spectrum and then stack them accordingly. The second, and the one that will be used in the thesis project is to illuminate the sample with just a narrow waveband and then take an intensity image and stack each subsequent waveband to create the hypercube.

3-1-2 Resolution

Two resolutions are to be considered, namely the spatial and spectral resolution. Spatial resolution is easily interpreted as the number of pixels used to capture an image and spectral resolution as the number wavebands used to break up the electromagnetic spectrum. Obviously higher resolution holds more information and is thus preferable, but the limitations of the microscope and computational power for a given application have to be taken into account.

3-1-3 Bit-depth

Finally, the bit depth is to be taken into account. It refers to the amount of intensity levels that can be captured for a given pixel and waveband. RGB is generally captured in 8-bit, but in HI higher bit-depths are preferable. Higher bit-depth is generally used to achieve a higher signal to noise ratio but also increases the computational power required to process the data. The bit-depth is furthermore limited by the sensor.

3-2 Spectral Preprocessing

Since many hyperspectral processing techniques are only interested in spectral information and not in spatial information the hypercube is generally unstacked resulting in a 2D matrix of $N_x \times N_y$ by L, where N_x is the amount of pixels in the x-direction, N_y is the amount of pixels in the y-direction and L is the number of wavebands. The way the image is unstacked is not important as long as the restacking later on is done accordingly. After the hypercube is unstacked the new data matrix, denoted by X, often undergoes preprocessing. This process tries to remove known disturbances and enhance features that are important for analysis. The preprocessing in this research will consist of two steps. the first is a step where some processes are applied to the data which aim to correct for flaws in the setup and/or equipment. In the second step we aim to remove some of the noise in the image.

3-2-1 Correction for the Setup

The first disturbances that need to be removed are the ones that are inherent to the setup. The most important are the uneven illumination and zero-input pixel values. The former disturbance is caused by the light source generally not producing the same intensity in all wavelengths, often lowering in intensity towards the maximum and minimum wavelength it can produce. This problem is easily solved by measuring the average intensity in the sensor at all wavelengths without a sample present. These values can then be used to determine a multiplier which compensates for this uneven illumination. However, this does not suddenly create new information in the less powerful wavelengths and more even illumination in the setup is always preferable. The latter disturbance is caused by the sensor. By putting the lens cap on the camera and taking multiple images it can be determined whether some pixels tend have a nonzero value given zero input that cannot be accounted for by noise. This is easily compensated for by determining which pixel present this behaviour, averaging its value and subtracting this bias from the data.

3-2-2 Noise Removal

The next disturbance that is to be minimised is the the noise in the image. The subject of denoising was already touched upon in 2-6 in the form of smoothing. However, using spatial filtering as is the case in smoothing also alters the spectra of the individual pixels in the process. Since retaining as much information in these spectra is essential to the workings of endmember extraction, these spatial filtering techniques are best avoided in HI. In this section we'll be looking at two different filtering techniques. The first filters in the spectral

domain instead of the spatial domain, thus retaining more spectral information, and the second technique uses a three dimensional kernel to convolve the image.

Savitzky Golay Filtering

The first possible approach is the Savitzky Golay filter[32]. In this method the spectrum is convolved with a window of a chosen length, where each subset of datapoints is fitted with a polynomial of an order less than the window size, minimizing the least squares error. The output spectrum will consist of the respective value on the polynomial of the central points of each subset. A polynomial of an order of one less than the window length will result in a perfect recreation, thus not changing the data, while a polynomial of a much smaller order will result in very aggressive smoothing. The Savitzky Golay filter is popular as it is less likely to distort the signal trend. However it has no edge preserving characteristics like the bilateral.

3D Gaussian Filtering

The second type of filtering that is considered for this problem is 3-dimensional gaussian filtering. Due to the relative simplicity of the gaussian filter it is possible to apply this in a 3-dimensional fashion without making it too computationally heavy. This way both spectral and spatial information will be taken into account in the denoising process. The three dimensional kernel is given by

$$K = \frac{1}{(\sqrt{2\pi}\sigma)^3} \exp\left(-\left(\frac{x^2+y^2+z^2}{2\sigma^2}\right)^p\right) \tag{3-1}$$

convolving this kernel with the hypercube will likely result in a denoised hyperspectral image which has clearer waveband images compared to the savitzky golay method, which could also result in clearer abundance maps later on in the process, but might be less successful in preserving information in the spectral domain. This could in turn lead to lesser performance in the endmember extraction process.

3-3 Virtual dimensionallity

As all unwanted effects have now been removed we can start looking towards the process of estimating the source signals, the endmember extraction. However, many endmember extraction algorithms require the amount of signal sources to be known a priori. Since this is not the case in our intended use, this value will need to be estimated. This is where the concept of virtual dimensionallity (VD) comes into play. VD is defined as the number of spectrally distinct signatures in a hyperspectral image. Multiple methods have been developed to estimate this value, many originating from different fields. A large sum of these make one of two assumptions (or both) that are might not work in our case. The first being that the signal sources to have a significant influence on the eigenvalues, which is not always the case when the target is low in occurance. The second being that the noise is gaussian white, which is generally not true in hyperspectral data. Considering this, one method seems particularly

of interest for this research, namely HySime[33]. This method does not require any input from the user and also immediately gives a subspace to project the data onto for optimal dimensionallity reduction in the sense of the mean square error. In order to do this the signal and noise correlation matrices are estimated and then a subset of eigenvectors is selected that minimizes the mean square error between the signal subspace and the data. First the noise needs to be estimated. This is done by assuming that each data sample vector can be represented as a linear mixture of the other data sample vectors. This assumption holds for most data vectors as long as the total amount of data vectors is significantly larger than the amount of signal sources, which is generally the case in HI. Let $\boldsymbol{X} = [\boldsymbol{r_1}, \boldsymbol{r_2}, \ldots, \boldsymbol{r_N}]$ be the data matrix of L by N where L is the amount of spectral bands and N the total amount of pixels. Now let $\boldsymbol{Z}$ be the transverse of $\boldsymbol{X}$ its transverse and $\boldsymbol{Z}_{\delta_i}$ be $\boldsymbol{Z}$ with the ith column missing. Then

$$\boldsymbol{z}_1 = \boldsymbol{Z}_{\delta_i}\boldsymbol{\beta}_i + \boldsymbol{\xi}_i$$

where z_i is the ith column of Z, β is the regression vector and ξ_i is the modelling error. The regression vector can now be estimated in in the least squares sense by

$$\hat{\boldsymbol{\beta}}_i = (\boldsymbol{Z}_{\delta_i}^T\boldsymbol{Z}_{\delta_i})^{-1}\boldsymbol{Z}_{\delta_i}^T\boldsymbol{z}_i$$

and the noise is then estimated by

$$\hat{\boldsymbol{\xi}}_i = \boldsymbol{z}_i - \boldsymbol{Z}_{\delta_i}\hat{\boldsymbol{\beta}}_i$$

Exploiting the relation between $(\boldsymbol{Z}_{\delta_i}^T\boldsymbol{Z}_{\delta_i})^{-1}\boldsymbol{Z}_{\delta_i}^T$ and $\boldsymbol{Z}$ the noise can be estimated for every data vector with little computational load. For the complete derivation the reader is referred to the referred article. Let $\hat{\boldsymbol{R}} = (\boldsymbol{Z}^T\boldsymbol{Z})$ and $\boldsymbol{R}'$ be its inverse. Now let $[\hat{\boldsymbol{R}}]_{\delta_i,\delta_i}$ denote the $\boldsymbol{R}$ matrix with the ith column and row deleted, $[\hat{\boldsymbol{R}}]_{i,\delta_i}$ the ith row of $[\hat{\boldsymbol{R}}]_{\delta_i,\delta_i}$, $[\hat{\boldsymbol{R}}]_{\delta_i,i}$ the ith column of $[\hat{\boldsymbol{R}}]_{\delta_i,\delta_i}$ and finally $[\hat{\boldsymbol{R}}]_{i,i}$ be the value of the ith row and ith column. Having the same rules apply to $\boldsymbol{R}'$ then $\boldsymbol{\beta}$ can be estimated in the following way

$$\hat{\boldsymbol{\beta}}_i = ([\hat{\boldsymbol{R}}]_{\delta_i,\delta_i} - [\hat{\boldsymbol{R}}']_{\delta_i,i}[\hat{\boldsymbol{R}}']_{i,\delta_i}/[\hat{\boldsymbol{R}}']_{i,i})[\hat{\boldsymbol{R}}']_{\delta_i,i}$$

In this case $\boldsymbol{R}$ and $\boldsymbol{R}'$ are determined outside of the loop determining β and ξ thus lowering the computational complexity. Knowing $\hat{\boldsymbol{\xi}}$ the sample, noise and signal correlation matrix can be estimated by

$$\begin{aligned}
\hat{\boldsymbol{R}}_y &= (\boldsymbol{X}\boldsymbol{X}^T)/N \\
\hat{\boldsymbol{R}}_n &= \frac{\Sigma_{i=1}^{L}(\hat{\boldsymbol{\xi}}_i\hat{\boldsymbol{\xi}}_i^T)}{N} \\
\hat{\boldsymbol{R}}_x &= \frac{\Sigma_{i=1}^{L}((\boldsymbol{r}_i - \hat{\boldsymbol{\xi}}_i)(\boldsymbol{r}_i - \hat{\boldsymbol{\xi}}_i)^T)}{N}
\end{aligned}$$

Let $\boldsymbol{E = e_1, \ldots, e_L}$ be the eigenvectors of $\hat{\boldsymbol{R}}_x$. When projecting onto a subset of E the mean square error is determined by the projection error power (parts of the signal not projected onto the subset) and the noise power (noise projected onto the subset). The minimization problem can be solved by determining

$$\delta_i = -\boldsymbol{e_i^T \hat{R}_y e_i} + \boldsymbol{2 e_i^T \hat{R}_n e_i}$$

and sorting by ascending order whilst remembering the permutation. The estimated number of endmembers is then equal to the number of terms $\delta_i \leq 0$ and the the corresponding eigenvectors span the signal subspace.

3-4 Dimensionallity Reduction

Given an estimate of the amount of signal sources, the next step of extracting these signal sources can be done in a number of ways, most belonging to some sort of endmember extraction. However, most of these methods require the dimensionallity to be reduced to a certain dimension, generally one less than the VD. Methods can generally be categorized as either feature selection or feature extraction. In case of HI feature selection would be selecting the wavebands that are statistically most interesting by some criterion. Feature extraction is the more interesting of the two. It performs a transform where each new feature is a combination of the original wavebands. Again the statistically most interesting of the new features are selected. Note that for some endmember extraction methods discussed in the following section dimensionallity reduction (DR) is included, whilst others require it to be performed prior, leaving multiple DR options to choose from. This section will discuss principle component analysis and independent component analysis.

3-4-1 Principle Component Analysis

The most well known method in feature extraction is principle component analysis (PCA) which is an eigenvalue based transform where combinations of wavebands that maximize the variance are found. It does so by finding the eigenvectors that belong to the biggest eigenvalues of the covariance matrix. Consider the data matrix $\boldsymbol{X} = [\boldsymbol{r}_1 \boldsymbol{r}_2 ... \boldsymbol{r}_N]$ where $\boldsymbol{r}_i$ is a L-dimensional spectrum of a pixel and N the total amount of pixels. The mean of the data set is given by $\boldsymbol{\mu} = (1/N)\Sigma_{i=1}^{N} \boldsymbol{r}_i$. The Covariance matrix can now be determined by

$$\boldsymbol{K} = (1/N)\Sigma_{i=1}^{N} (\boldsymbol{r}_i - \boldsymbol{\mu})(\boldsymbol{r}_i - \boldsymbol{\mu})^T$$

Which can be rewritten as

$$\boldsymbol{K = \Lambda D_\sigma \Lambda^T}$$

where $\boldsymbol{D}_\sigma = diag(\sigma_1, \sigma_2, ..., \sigma_N)$, $sigma_i$ being the ith eigenvalue of the covariance matrix, and $\boldsymbol{\Lambda} = [v_1 v_2 ... v_N]$ where v_i are the corresponding eigenvectors. The covariance and eigenvector

matrix are then ordered in descending order of eigenvalue denoted by $\boldsymbol{D}'_\sigma$ and $\boldsymbol{\Lambda}'$. The data is then transformed

$$\hat{\boldsymbol{X}} = \boldsymbol{\Lambda}^{\boldsymbol{T}} \boldsymbol{X}$$

each row now being an principal component. Dimensionallity reduction is achieved by only retaining a certain amount of the principle components corresponding to the biggest eigenvalues. Normally, the amount of principle components to retain is often chosen by looking for a big jump in the size of the eigenvalues. Only those components corresponding to the eigenvalues prior to the gap (in descending order) are then retained. However in case of endmember extraction the dimensionallity is often simply reduced to one less than the VD.

3-4-2 Independent Component Analysis

Up until now only the variance has been used in order to find the components. However, the components with the highest variance are not necessarily the ones of interest. An issue with second order statistics based component analysis methods is that they generally have problems finding many subtle substances due to their small amounts and little contribution to the second order statistics. This is where higher order statistics based component analysis comes in. One of the most prominent, though its use in dimensionallity reduction [34] is quite new, is independent component analysis (ICA), first introduced by Jutten and Herault [35]. This method uses a combination of higher order statistics such as skewness and kurtosis to measure statistical independence. In order to accomplice this the first and second order statistics (mean and variance) need to be removed. This is done by a technique called sphering. Let $\boldsymbol{X} = [\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_N]$ again be the data matrix. The sample mean is first removed by

$$\tilde{\boldsymbol{X}} = \boldsymbol{X} - \boldsymbol{\mu} \times \mathbf{1}^T$$

where

$$\boldsymbol{\mu} = (\frac{1}{N}) \Sigma_{i=1}^{N} \boldsymbol{r_i}$$

and $\mathbf{1}$ is a column vector with all ones as components. Then the variance is to be removed. Let $\boldsymbol{K}_{\tilde{\boldsymbol{X}}} = (\frac{1}{N}) \tilde{\boldsymbol{X}} \tilde{\boldsymbol{X}^T}$ be the sample covariance matrix of $\tilde{\boldsymbol{X}}$ and $\{\lambda_l\}_{l=1}^{L}$ and $\{\boldsymbol{v}_l\}_{l=1}^{L}$ to be its eigenvalues and eigenvectors respectively. Suppose $\boldsymbol{\Lambda}$ to be the eigenvector matrix, then

$$\boldsymbol{\Lambda}^{\boldsymbol{T}} \boldsymbol{K}_{\tilde{\boldsymbol{X}}} \boldsymbol{\Lambda} = \boldsymbol{D}_\lambda$$

where $\boldsymbol{D}_\lambda = diag(\lambda_1, \lambda_2, \ldots, \lambda_L)$. From this we conclude that

$$(\boldsymbol{D}_\lambda)^{-1/2} \boldsymbol{\Lambda}^{\boldsymbol{T}} \boldsymbol{K}_{\tilde{\boldsymbol{X}}} \boldsymbol{\Lambda} (\boldsymbol{D}_\lambda)^{-1/2} = \boldsymbol{I}$$

and the sphering matrix $\boldsymbol{M} = \boldsymbol{\Lambda}(\boldsymbol{D})^{-1/2}$ is obtained. The data can now be transformed

$$\boldsymbol{X}_{sph} = \boldsymbol{M}\boldsymbol{X}$$

ICA can be implemented in many ways. Initially many methods suggested to find a vector of norm one that yields maximum or minimum kurtosis, the fourth order statistic. Kurtosis is defined as

$$\mathbf{kurt}(v) = E\{\boldsymbol{v}^4\} - 3(E\{\boldsymbol{v}^2\})^2$$

for random variable $\boldsymbol{v}$. Let $||w||$ be bounded by 1. There exists a linear combination of sphered observations $\boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{X}$ which yields maximum kurtosis. An objective function is determined as

$$\begin{aligned}\mathbf{kurt}(\boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{X}_{sph}) &= E\{(\boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{X}_{sph})^4\} - 3(E\{\boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{X}_{sph})^2\})^2 \\ &= E\{(\boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{X}_{sph})^4\} - 3||\boldsymbol{w}||^2\end{aligned}$$

A well known algorithm is the FastICA [36], in which a very fast iteration is utilized derived from this equation. It consist of just a few steps

- Take a random initial vector $w(0)$ of norm 1. let $k = 1$
- Let $\boldsymbol{w}(k) = E[\boldsymbol{X}_{sph}(\boldsymbol{w}(k-1)^T\boldsymbol{X}_{sph})^3] - 3\boldsymbol{w}(k-1)$. The estimation can be estimated using a large sample of $\boldsymbol{X}_{sph}$ vectors (say 1000 points)
- divide $\boldsymbol{w}(k)$ by its norm.
- If $|\boldsymbol{w}(k)^T\boldsymbol{w}(k-1)|$ is not close enough to one, let $k = k + 1$ and go back to step 2. Otherwise output vector $\boldsymbol{w}(k)$

If multiple vectors need to be found simply iteratively use orthogonal projection to remove the previously found vector from the data set.

- let $k = 1$
- let $k = k + 1$
- Use the FastICA algorithm to find w_k
- apply the orthogonal projection matrix $P_{w_k} = \boldsymbol{I} - \boldsymbol{w_k}((\boldsymbol{w_k})^{\boldsymbol{T}}\boldsymbol{w_k})^{-\mathbf{1}}(\boldsymbol{w_k})^{\boldsymbol{T}}$ to the dataset $\boldsymbol{X}_{sph}^{-k} = \boldsymbol{P}_{\boldsymbol{w}_k}\boldsymbol{X}_{sph}^{-k+1}$.
- if k is the amount of previously determined dimension than stop, if not return to step 2.

Important to note is that ICA assumes there to be at most one gaussian source as the mean and variance is removed by the sphering. This means that these sources cannot be separated by ICA.

3-5 Endmember Extraction

Now that dimensionallity reduction is achieved the next step is to find the endmembers of the data as these are an estimation of the sources present in the data. An endmember is an idealised pure spectrum of a class. These endmembers can provide meaningful insight into the sample. In the ideal case a single endmember can be matched to the parasite. This endmember can then later be used to detect the parasite in a new sample. The field of endmember extraction methods is very extensive with a multitude of algorithms that have been found to be well performing in various different applications. A few of the most commonly used methods are discussed in this section. A big group of methods can be categorized as pure pixel based. They assume for each endmember there is at least one pure pixel in the image. Two methods are discussed that fall outside this classification. These try to find endmembers that best represent the data by some criterion.

3-5-1 Pure Pixel Index

The first endmember extraction method to be discussed is the pure pixel index (PPI). This method assumes endmembers to be present as a pure pixel in the data and is well known for its in the ENVI software. However, literature about the method is scarce and thus the version Chang [10] describes in his 2013 book is considered. This version is tested against the ENVI software to performs similarly. For PPI the dimensionallity of the the data is first reduced to the VD minus one. After this the endmembers are determined using the assumption that if a data point is an endmember it is likely to give a maximum or minimum orthogonal projection on a random vector. First k random vectors are created denoted by $\{\boldsymbol{s}_k\}_{i=1}^{k}$ and are called skewers. k has to be chosen sufficiently large in order to achieve good performance. When the skewers are created each data sample vector $\boldsymbol{r}_i$ is orthogonally projected onto each skewer by

$$P_{\boldsymbol{s}_k}(r) = \frac{\boldsymbol{s}_k \boldsymbol{s}_k^T}{\boldsymbol{s}_k^2} \boldsymbol{r}_i$$

The set of skewers onto which r produces either a maximal or minimal projection is called $S_{extrema}(r)$. The PPI count n_{PPI} is then determined by the number of skewers in $S_{extrema}(\boldsymbol{r}_i)$. A threshold t is then chosen on the count n_{PPI} and finally all the sample vectors with $n_{PPI} \geq t$ are extracted as endmembers. An example is given in 3-3. Three random skewers are created. $\boldsymbol{e}_1$ is shown to have 3 maxima/minima, $\boldsymbol{e}_2$ has 2 and $\boldsymbol{e}_3$ has 1. The grey circle $\boldsymbol{x}$ is also shown to have an PPI count of 1 as it projects on the same minimum of skewer 1 as $\boldsymbol{e}_2$. However, when calculating the volumes of the triangles spanned by three out of four possible endmembers, $\{\boldsymbol{e_1}, \boldsymbol{e_2}, \boldsymbol{e_3}\}$ produces the maximum volume. Thus $\{\boldsymbol{e_1}, \boldsymbol{e_2}, \boldsymbol{e_3}\}$ are likely the desired endmembers.

This technique has several drawbacks. Since the algorithm is not iterative, but determines all the endmembers at once while being dependant on the random initialization of the skewers, each run of the algorithm can give a different result. Furthermore, the results are very dependant on the choice of K and t. A bad choice could result in certain important signatures missing in the endmembers or having multiple endmembers represent the same spectral signature.

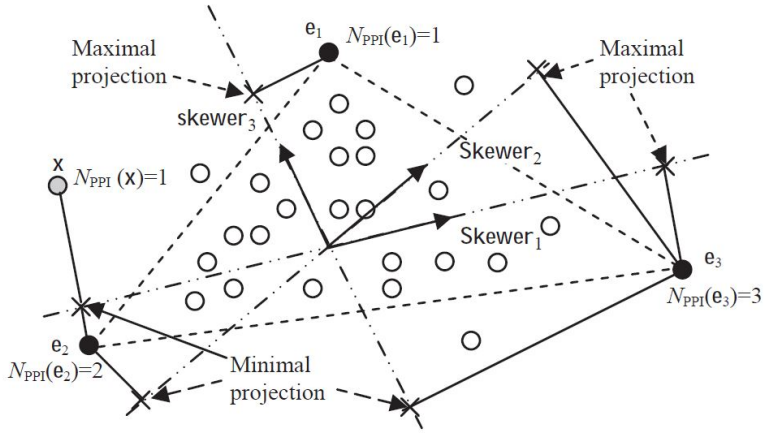


Figure 3-3: Illustration of the PPI endmember extraction algorithm displaying several data points (open circles), skewers (arrows) and endmembers (dark circles) [10].

3-5-2 N-FINDR

Another widely used technique is the N-FINDR, developed by Winter [37]. This technique is also of the former kind assuming endmembers to be present as a pure pixel in the data. It assumes that in N dimensions the so called N-volume contained by a simplex with the endmembers on its vertices is the largest possible simplex in terms of volume. Once again, dimensionallity reduction is performed prior by one of the techniques described in section 3. After determining the VD denoted by p the data is transformed to a dimensionallity of $p-1$. The volume of the simplex with any p data sample vectors $\boldsymbol{e_1}, \boldsymbol{e_2}, \ldots, \boldsymbol{e_p}$ denoted by $S(\boldsymbol{e_1}, \boldsymbol{e_2}, \ldots, \boldsymbol{e_p})$ is given by

$$V(\boldsymbol{e_1}, \boldsymbol{e_2}, \ldots, \boldsymbol{e_p}) = \frac{\left|\det \begin{bmatrix} 1 & 1 & \ldots & 1 \\ \boldsymbol{e_1} & \boldsymbol{e_2} & \ldots & \boldsymbol{e_p} \end{bmatrix}\right|}{(p-1)!}$$

Then an exhaustive search is performed to find the combination of data sample vectors for which the volume of the simplex is maximal

$$\{\boldsymbol{e_1^*}, \boldsymbol{e_2^*}, \ldots, \boldsymbol{e_p^*}\} = \arg\{max_{\boldsymbol{e_1}, \boldsymbol{e_2}, \ldots, \boldsymbol{e_p}} V(\boldsymbol{e_1}, \boldsymbol{e_2}, \ldots, \boldsymbol{e_p})\}$$

where $\{\boldsymbol{e_1^*}, \boldsymbol{e_2^*}, \ldots, \boldsymbol{e_p^*}\}$ are the desired endmembers. Unlike PPI this method does not require the user to choose certain parameters, yet does still suffer from the drawback that all endmembers need to be determined simultaneously. The exhaustive search can be computationally quite heavy depending on the amount of datapoints and endmembers. However, a variation on the N-FINDR algorithm has been developed called the simplex growing algorithm [38]. In this algorithm the endmembers are determined sequentially instead of all at once. The process is described as followed

1. Initialization: A random endmember is selected from the data and let $k = 0$.
2. at $k \geq 0$ the original L dimesional data is transformed to k dimensions by dimensionallity reduction method. Then the volume of the $k+1$-dimensional simplex is determined for each data sample vector r

$$V(\boldsymbol{e^{(0)}}, \ldots, \boldsymbol{e^{(k)}}, \boldsymbol{r}) = \frac{\left| \det \begin{bmatrix} 1 & 1 & \ldots & 1 & 1 \\ \boldsymbol{e^{(0)}} & \boldsymbol{e^{(1)}} & \ldots & \boldsymbol{e^{(k)}} & r \end{bmatrix} \right|}{(k)!}$$

3. The new endmember is the data sample vector which corresponds to the largest volume

$$\boldsymbol{e}^{(k+1)} = \arg\{max_r[V(\boldsymbol{e^{(0)}}, \ldots, \boldsymbol{e^{(k)}}, \boldsymbol{r})]\}$$

4. If $k \leq p-1$ then $k \rightarrow k+1$ and go to step 2, otherwise $\{\boldsymbol{e^{(1)}}, \boldsymbol{e^{(2)}}, \ldots, \boldsymbol{e^{(p)}}\}$ is the final set of endmembers.

As this version of the algorithm only determines one endmember at a time it demands much less computational power. Furthermore, if the amount p is increased the algorithm can continue using the previously found endmembers. This could be useful if a certain set of endmembers in the data is already known. A drawback of the algorithm is that its performance is very dependant on the initial randomly selected endmember.

3-5-3 SISAL

Another simplex based approach is to minimize the volume of a simplex that encloses all the data instead of inflating a simplex inside the data as is the case with N-FINDR, a concept introduced by Craig [39]. However, since this algorithm is computationally very heavy multiple publications have be done to increase performance by reformulating the optimization problem. The Algorithm discussed here is the simplex identification via split augmented Lagrangian (SISAL) [40], which proposes to solve the optimization problem by using soft constraints and then use a sequence of augmented Lagrangian optimizations. The concept is based on the assumptions that the vertices of the simplex that encloses the data and is minimal in volume coincide with the endmembers. Since it is too computationally heavy to do an exhaustive search an optimization algorithm is applied. Let $\boldsymbol{X} = [\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_N] \in R^{LxN}$ be the data matrix. Assuming each data sample vector is well approximated by a linear mixing model results in

$$\boldsymbol{X} = \boldsymbol{MA} + \boldsymbol{N}$$

where $\boldsymbol{M} = [\boldsymbol{m}_1, \boldsymbol{m}_2, \ldots, \boldsymbol{m}_p]$ is the the mixing matrix containing the endmembers and A is the abundance matrix containing all the abundance vectors. The matrices M and A are then

to be determined by fitting a minimum volume simplex that encapsulates all the data. In the SISAL method this optimization problem is rewritten to

$$\begin{aligned} Q* &= \arg\min(-log|\det(Q)| + \lambda\|QX\| \\ )\text{s.t.:} \quad & \mathbf{1}_p^T Q = a^T \end{aligned}$$

where $\boldsymbol{Q} \equiv \boldsymbol{M}^{-1}$, p is the amount of endmembers and $a = \mathbf{1}_n \boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}$. Here $\lambda\|\boldsymbol{QY}\|$ represents a soft constraint on negative abundances and $\mathbf{1}_p^T Q = \boldsymbol{a}^T$ represents a hard sum-to-one constraint on the abundances. The Lagrangian of this problem is given by

$$\mathcal{L}(\boldsymbol{q}, \boldsymbol{z}, \boldsymbol{d}, \tau) \equiv \boldsymbol{g}^T\boldsymbol{q} + \mu\|\boldsymbol{q} - \boldsymbol{q_k}\|^2 + \lambda\|\boldsymbol{z}\| + \tau\|\boldsymbol{Cq} - \boldsymbol{z} - \boldsymbol{d}\|^2 + c$$

where q is column vector containing the columns of $\boldsymbol{Q}$, $\boldsymbol{z} = \boldsymbol{Cq}$, $\boldsymbol{d} = -\frac{\boldsymbol{Bq}}{2\tau}$, where $\boldsymbol{C} = (\boldsymbol{X}^T \otimes \boldsymbol{I})$ and $\boldsymbol{B} = (\boldsymbol{I} \otimes \mathbf{1}_{\boldsymbol{p}}^{\boldsymbol{T}})$. The term $\mu\|\boldsymbol{q} - \boldsymbol{q_k}\|^2$ prevents unbounded growth (μ being a regularizing parameter) and finally $\tau\|\boldsymbol{Cq} - \boldsymbol{z} - \boldsymbol{d}\|^2$ can be interpreted as the constraints, now all soft. The problem can now be solved iteratively,

1. set $t = 0$ and choose ($\boldsymbol{q}_0$, $\boldsymbol{z}_0$, $\boldsymbol{a}_0$ and $t > 0$
2. repeat
3. $(\boldsymbol{q}_{t+1}, \boldsymbol{z}_{t+1}) \in \arg\min_{\boldsymbol{z}} \mathcal{L}(q, z, d, \tau)$
   s.t. $\boldsymbol{Bq} = \boldsymbol{a}$
4. $\boldsymbol{d}_{t+1} = \boldsymbol{d}_t - (\boldsymbol{A}\boldsymbol{q}_{\boldsymbol{t+1}} - \boldsymbol{z}_{\boldsymbol{t+1}})$
5. $t = t + 1$
6. until stopping criterion is met

where the starting values $\boldsymbol{q}_0$ and $\boldsymbol{z}_0$ are derived from another pure pixel based method, namely VCA. Due to the soft constraints and implementation of augmented lagrangian, this method is much faster than other minimum volume simplex methods and has been shown to perform similarly in case of lower amounts of endmembers, as is the case in this application. The method is still slower than the other methods discussed in this research, but not unreasonably so. It could however provide significantly better results in some circumstances.

3-5-4 Statistics based

However, in the case that the data is highly mixed the endmembers found by previous methods might be much smaller than the true endmembers as shown in figure 3-4. In this case a statistics based approach might provide a solution. Using the components found by higher order dimensionallity reduction such as ICA for endmembers extraction might provide better results. A method such as FastICA is used to determine p components. The maximal projection of the data vectors onto each component is determined. The set of spectral signatures corresponding to these maximal projections, denoted by $\{\boldsymbol{e}_j\}_{j=1}^p$, is the desired set of endmembers.

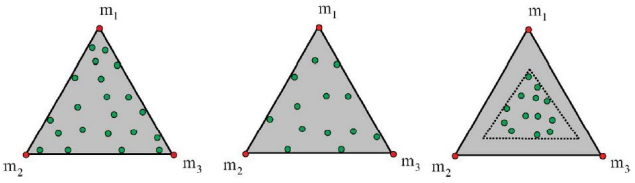


Figure 3-4: Minimum volume simplex of three data sets. In the left and middle case the endmembers can be found by a minimum volume simplex algorithm. In the right case this is not possible. This corresponds to heavily mixed data. [11]

3-6 Spectral Unmixing

We have now found the endmembers that most likely make up the data. These should represent spectral signatures of substances in the sample. However, since the endmembers are found in an unsupervised manner, they need to be analysed. Ideally, one endmember can be matched to a type of parasite, as this could make the parasite easily detectable in new future samples. One way to analyse these endmembers is to use the endmembers to unmix the hyperspectral images and look at there respective abundance images. The unmixing of the hyperspectral image can be done by Fully Constrained Least Squares (FCLS) unmixing. In this method the well known least squares regression, which would result in the abundances being given by

$$\hat{\alpha}_{LS} = (\boldsymbol{M^T M})^{-1} \boldsymbol{M^T r}$$

is altered to include two constrains relevant to our problem, namely that the abundance of a signature cannot be negative and that the abundances must sum to one in each pixel. Taking these constraints into account the regression problem becomes a lot more complex. Fully constrained least squares linear unmixing (FCLSLU), as proposed by [41] was first method able to find a solution to this optimization problem. First the solution to the nonnegativity constrained least squares (NCLS) problem is derived

$$\begin{aligned} min_\alpha\{(r - M\alpha)^T(r - M\alpha)\} \\ \text{s.t.} \quad \alpha \geq 0 \end{aligned}$$

Since the constraint is not an equality constraint Lagrange multipliers cannot be used to solve the optimization. However, by introducing an unknown vector $c = (c_1, c_2, \ldots, c_p)^T$ a Lagrange multiplier can be introduced in the following way.

$$J = \frac{1}{2}\{(r - M\alpha)^T(r - M\alpha)\} + \lambda(\alpha - c)$$

with $\alpha = c$ and

$$\|\frac{\partial J}{\partial \alpha}\|_{\hat{\alpha}_{NCLS}} = 0 \rightarrow \boldsymbol{M^T M}\hat{\alpha}_{NCLS} - \boldsymbol{M^T r} + \lambda = 0$$

which leads to the equations

$$\hat{\alpha}_{NCLS} = \boldsymbol{(MM^{T-1}M^T r - (M^T M)^{-1}}\lambda \tag{3-2}$$
$$= \alpha_{LS} - (M^T M)^{-1}\lambda \tag{3-3}$$

and

$$\lambda = \boldsymbol{M^T(r - M}\hat{\alpha}_{NCLS})$$

Two index sets are formed. P consists of all the indices corresponding to all the positive components of α_{LS} and R consists of all the indices corresponding to all the negative or zero components of α_{LS}. An optimal solution must then satisfy the following Kuhn-Tucker conditions

$$\lambda_j = 0 \quad j \in P$$
$$\lambda_j < 0 \quad j \in R$$

The following iterative process can then be used to find α_{NCLS}

1. set $P^{(0)} = \{0, 1, \ldots, p$, $R^{(0)}$ empty and $k = 0$
2. compute $\hat{\alpha}_{LS}$ using 3-2 and let $\hat{\alpha}^{(0)}_{NCLS} = \hat{\alpha}_{LS}$
3. If all components of $\hat{\alpha}^{(0)}_{NCLS}$ are negative the algorithm is terminated. If not, continue.
4. let $k \leftarrow k + 1$
5. All indices corresponding to negative values in $\hat{\alpha}^{(k-1)}_{NCLA}$ are moved from $P^{(k-1)}$ to $R^{(k-1)}$ resulting in $P^{(k)}$ and $R^{(k)}$. Let $S^{(k)}$ be equal to $R^{(k)}$
6. Let $\hat{\alpha}_{R^{(k)}}$ be the vector consisting of all components of $\hat{\alpha}_{LS}$ corresponding to $R^{(k)}$
7. A steering matrix $\phi^{(k)}_{\alpha}$ is then formed by deleting all the rows and columns of $\boldsymbol{(MM)}^{T-1}$ that correspond to $P^{(k)}$.
8. Let the new lagrange multiplier be given by $\lambda^{(k)} = (\phi^{(k)}_{\alpha})^{-1}\hat{\alpha}_{R^{(k)}}$. If all entries are negative, go to step 13, otherwise continue.
9. Determine the maximum lagrange multiplier λ_{max} and move its index to $P(k)$.

10. Another matrix $\psi_{\lambda}^{(k)}$ is formed by deleting all the columns from $(\boldsymbol{MM})^{T-1}$ that correspond to $P^{(k)}$.

11. Let $\hat{\alpha}_{S^{(k)}} = \hat{\alpha}_{LS} - \psi_{\lambda}^{(k)}\lambda^{(k)}$.

12. If any components of $\hat{\alpha}_{S^{(k)}}$ are negative, their corresponding indices are moved from $P^{(k)}$ to R. go to step 6

13. Another matrix $\psi_{\lambda}^{(k)}$ is formed by deleting all the columns from $(\boldsymbol{MM})^{T-1}$ that correspond to $P^{(k)}$.

14. Let $\hat{\alpha}_{NCLS}^{(k)} = \hat{\alpha}_{LS} - \psi_{\lambda}^{(k)}\lambda^{(k)}$. go to step 3.

Now the sum to one constraint has to be implemented. Luckily, this is easy to implement in the previously discussed framework by M by

$$N = \begin{bmatrix} \delta M \\ 1^T \end{bmatrix}$$

where $1 = (1, 1, \ldots, 1)^T$ of length p. Furthermore r is replaced by

$$s = \begin{bmatrix} \delta r \\ 1 \end{bmatrix}$$

where δ controls the impact of the sum-to-one constraint on the iteration process.

3-7 Semi-Supervised Spectral Signature Estimation

Of course, in the case of giemsa stained samples, it is easy to verify whether the algorithm worked as even in many of the waveband images the parasites are already visible. However, in case of the unstained samples, they are not. This makes it much harder to verify whether the algorithm worked properly. It is possible that the abundance maps derived from the hyperspectral image of the unstained sample clearly show the parasites, but this might be an overly optimistic expectation. However, if the algorithm works in on the stained sample, but does not work as hoped in the unstained case, there is still one more possibility that can be explored. Namely, to image the same part of a sample in both the stained and unstained condition. If the cells and parasites remain in the same position during the staining process, the images could be aligned. This would make it possible to use the abundance maps of the stained and unstained condition to derive the spectral signatures of the unstained image. Using the hyperspectral data of the unstained image and the abundance map of the stained image least squares regression could be used to determine these sought after signatures.

3-8 Target Detection

If the matching of endmembers to the parasites, RBC's and the like was successful, the next challenge is to be able to use these signatures to be able to detect the parasites in new samples. The previously discussed FCLS method could again be used for this, but is computationally very heavy and thus not suited for quick diagnostic methods. Luckily, there are many computationally lighter possibilities to tackle this so called target detection. The choice is often very dependant on the amount of information that is known a priori. In some cases it could only be necessary to just look at how similar the spectrum of a given pixel is to a reference spectrum and other cases might require the use of all the known endmembers to create a more reliable abundance map. This section will discuss some of the possibilities.

3-8-1 Similarity Measures

The first possibility is to just look at how similar the pixel spectrum is to a given reference spectrum. Two measures will be discussed in this report, euclidean distance and spectral angle mapper. They are often used as a computationally lighter way to inspect new samples.

Euclidean Distance

A widely used metric is the euclidean distance (ED) and is simply the distance between two points in euclidean space. In HI it can be used to determine the distance between two spectral signature s_i and s_j in the following way

$$ED(\boldsymbol{s_i}, \boldsymbol{s_j}) = ||\boldsymbol{s_i} - \boldsymbol{s_j}|| = \sqrt{\Sigma_{l=1}^{L}(\boldsymbol{s_{il}} - \boldsymbol{s_{jl}})^2}$$

Generally, the more similar the signatures the smaller the euclidean distance. However, when the illumination levels differ greatly between the two pixels, two very similar pixels could still have a significant euclidean distance. In such an instance it is often preferable to use Spectral angle mapping.

Spectral Angle Mapper

Spectral Angle Mapping (SAM) is a very popular technique in HI as it is not dependant on the intensity of the spectrum unlike euclidean distance. Two similar spectra but with very different illumination levels will therefore still have a small spectral angle. The spectral angle mapper determines the angle between two spectral signatures s_i and s_j in the following way

$$SAM(\boldsymbol{s}_i, \boldsymbol{s}_j) = \cos^{-1}\left(\frac{\langle \boldsymbol{s}_i, \boldsymbol{s}_j \rangle}{||\boldsymbol{s_i}||||\boldsymbol{s}_j||}\right)$$

where $\langle \boldsymbol{s}_i, \boldsymbol{s}_j \rangle = \Sigma_{i=1}^{L} \boldsymbol{s}_{il}\boldsymbol{s}_{jl}$, $||\boldsymbol{s}_i|| = (\Sigma_{l=1}^{L}(\boldsymbol{s}_{il})^2)^{\frac{1}{2}}$ and $||\boldsymbol{s}_j|| = (\Sigma_{l=1}^{L}(\boldsymbol{s}_{jl})^2)^{\frac{1}{2}}$. Once again the signatures are most similar when the SAM is small.

3-8-2 Constrained energy minimization

However, often the pixel will be a mixture of multiple of the endmembers. In such a case the spectrum will not be very similar to the reference spectrum. Thus, if the parasite spectrum is mixed with other spectra, euclidean distance and spectral angle mapper might not pick up on its presence. This is where constrained energy minimization (CEM) [42] provides a solution. A signal detector is designed by finding a finite impulse response linear filter that minimizes the filter output energy. Let the output of the filter be expressed by

$$y_i = (\boldsymbol{w})^T \boldsymbol{r_i} = (\boldsymbol{r}_i)\boldsymbol{w}$$

Where $\boldsymbol{w}$ is an L dimensional weighing vector. The average output is then given by

$$\frac{1}{N}\Sigma_{i=1}^{N} y_i^2 = \boldsymbol{w}^T \boldsymbol{R}\boldsymbol{w}$$

where $R = \frac{1}{N}\Sigma_{i=1}^{N} \boldsymbol{r}_i (r_i)^T$ is the auto correlation matrix. In order to minimize the output energy the following optimization problem is considered

$$\min_{w}\{\boldsymbol{w}^T \boldsymbol{R}\boldsymbol{w}\} \quad \text{s.t.} \quad \boldsymbol{d}^T\boldsymbol{w} = \boldsymbol{w}^T\boldsymbol{d} = 1$$

where d is the target signature. The optimal solution is derived to be

$$\boldsymbol{w}^{\text{CEM}} = \frac{\boldsymbol{R}^{-1}\boldsymbol{d}}{\boldsymbol{d}^T\boldsymbol{R}^{-1}\boldsymbol{d}}$$

The abundance can then be determined by

$$\alpha_d^{\text{CEM}}(r) = (\boldsymbol{w}^{\text{CEM}})^T\boldsymbol{r} = \frac{\boldsymbol{d}\boldsymbol{R}^{-1}r}{\boldsymbol{d}^T\boldsymbol{R}^{-1}\boldsymbol{d}}$$

Note that the result is not strictly speaking an abundance map, but an impulse response. The resulting image therefore does not have properties such as its values being between 0 and 1. However, it is very suitable to be used for visual inspection or for feature creation for classification purposes.

3-8-3 Orthogonal Subspace Projection

Finally, if all spectra present in the sample are known it would be ideal to be able to use this information. This is possible using orthogonal subspace projection. Using OSP for spectral unmixing, as first described by [43], the abundance of one desired target signature is sought

after, while the other signatures are suppressed. Let $\boldsymbol{d}$ be the desired target signatures and $\boldsymbol{U}$ be the undesired signature matrix. The undesired signatures are then removed from the signal prior to the detection of $\boldsymbol{d}$. Let each data vector be given by

$$\boldsymbol{r} = \boldsymbol{d}\alpha_p + \boldsymbol{U}\gamma + n$$

where α_p is the abundance value corresponding to the desired target signature, γ is the abundance vector corresponding to the undesired signatures and $\boldsymbol{n}$ is the noise. The undesired signatures are removed from the data by an orthogonal subspace projector

$$\boldsymbol{P}_{\boldsymbol{U}}^{\perp}\boldsymbol{r} = \boldsymbol{P}_{\boldsymbol{U}}^{\perp}\boldsymbol{d}\alpha_p + \boldsymbol{P}_{\boldsymbol{U}}^{\perp}\boldsymbol{n}$$

given

$$\boldsymbol{P}_{\boldsymbol{U}}^{\perp} = \boldsymbol{I} - \boldsymbol{U}\boldsymbol{U}$$

where U is the pseudo inverse of $\boldsymbol{U}$. Then SNR maximization is used as a criterion to find the optimal weighing vector w in

$$\boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{P}_{\boldsymbol{U}}^{\perp}\boldsymbol{r} = \boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{P}_{\boldsymbol{U}}^{\perp}\boldsymbol{d}\alpha_p + \boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{P}_{\boldsymbol{U}}^{\perp}\boldsymbol{n}$$

which is optimal by Schwartz inequallity

$$|\boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{P}_{\boldsymbol{U}}^{\perp}\boldsymbol{r}| = ||\boldsymbol{w}||||\boldsymbol{P}_{\boldsymbol{U}}^{\perp}\boldsymbol{d}||$$

which is optimal for $w = \kappa P_U^{\perp} d$ for some constant κ. Thus, the optimal detector becomes

$$\delta^{OSP}(r) = \kappa\boldsymbol{d}^{\boldsymbol{T}}\boldsymbol{p}_{\boldsymbol{U}}^{\perp}\boldsymbol{r}$$

OSP can be combined with CEM by implementing an orthogonal projection which removes some known undesired targets from the data prior to CEM. Let U be the matrix containing all the undesired signatures. The orthogonal projection matrix removing these signatures is then given by

$$\boldsymbol{P}_{\boldsymbol{U}}^{\perp} = \boldsymbol{I} - \boldsymbol{U}\boldsymbol{U}$$

Multiplying the data by this matrix prior to using CEM can greatly improve the results. However, in some cases, if the signature of the desired target is close to one of the signatures of the undesired targets this can result in significant deterioration of the signal detectabillity.

3-9 Classification

In the previous section the derived spectral signatures were used in order to create a detection image. The next step is to give a positive or negative diagnosis and determine the parasitemia. In order to do this first the RBC's have to be individually segmented from the image. When a RBC is segmented a set of features is to be derived in order to use as an input to the classifier. These features are derived from the detection image as these should clearly display the parasites and thus make for easier classification. This section will describe the proposed methods for each of methods.

3-9-1 RBC Segmentation

The subject of RBC segmentation in RGB images was already touched upon in the first chapter. The segmentation in the hyperspectral case is not very different other than that we start with a large hypercube. In order to apply regular RBC segmentation algorithms the data first has to be turned into a single intensity image. It is possible to use the detection image, but these generally display quite a lot of noise, making the segmentation process a lot harder. Therefore, the image displaying the first principal component of the hypercube is much better suited for this as it displays the signal with the highest variance and thus improves the signal-to-noise ratio. This image and the subsequent steps can be seen in figure 3-5. This image is then turned into a binary image by Otsu thresholding.

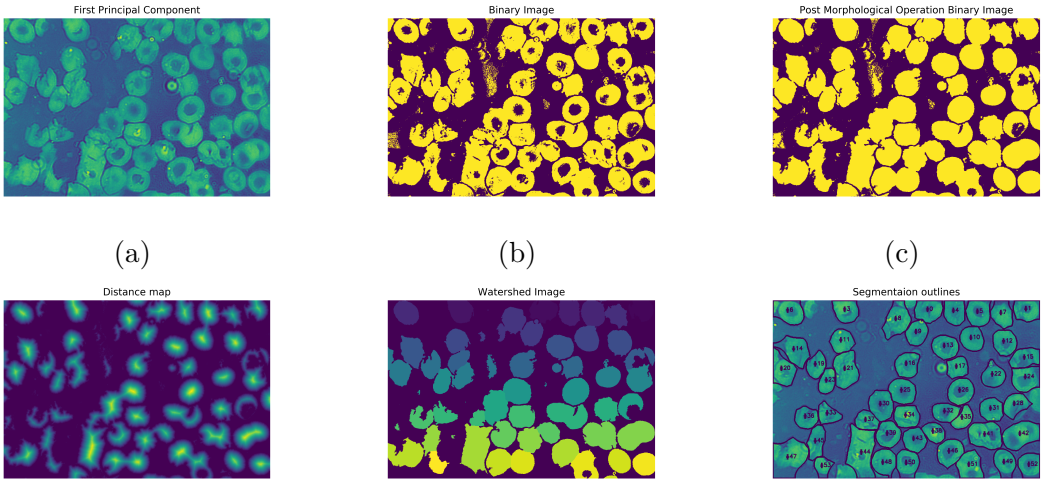


Figure 3-5: The RBC segmentation process: first principal component of the hyperspectral image (a), binary image after thresholding (b), binary image after morphological operations (c), distance map of the binary image (d), watershed image with different colors representing different masks (e), resulting masks drawn on original image (f)

Using various morphological operation including hole filling, opening and closing, the image should now display 1 values where the cells are and 0 values where there no cells are. It is now possible to segment the cells from the rest of the image. However, it is not yet possible

to segment clustered cells from each other. In order to do this a watershed algorithm is used. First, a distance map is created, where the values of the pixels represent the distance to the closest 0 value in the binary image. Local maxima are, with a minimum distance between them, are likely located in the middle of cells. A watershed algorithm starting in these local maxima is then able to separate small clusters. Since RBC segmentation is not the focus of this research this is considered sufficient.

3-9-2 Feature Extraction

Many modern day classification is done using neural networks and deep learning, where the whole detection image could be used as input to the classifier. However, since a new dataset had to be created for this research and the amounts of data required for these types of classifiers are much larger than was feasible in the timeframe, more traditional classifiers were considered. These classifiers do however require a set of features to be derived from the image to be used as inputs. A set of 10 features, some statistical in nature and some textural in nature, was decided on.

Histogram descriptors

An easy way to determine determine a feature is to derive a histogram from an intensity image containing the intensity values and their number of occurrences. These histograms can then be described by a few descriptors. The most obvious choices are mean value and variance, but many more are possible. Common choices include but are not limited to entropy and higher order statistics like skewness and kurtosis. Let $H(i)$ be the histogram that denotes the number of times the intensity i is present in an intensity image. The histogram is then normalized by

$$H'(i) = \frac{H(i)}{\Sigma_i H(i)} \tag{3-4}$$

and some statistical features are then given by

$$MV = \Sigma_i i H'(i) \tag{3-5}$$
$$VA = \Sigma_i (i - MV)^2 H'(i) \tag{3-6}$$
$$SK = \frac{\Sigma_i (i - MV)^3 H'(i)}{\sigma^3} \tag{3-7}$$
$$KU = \frac{\Sigma_i (i - MV)^4 H'(i)}{\sigma^4} \tag{3-8}$$

Haralick descriptors

Another popular choice is textural features. Introduced by Haralick [30], these feature try to capture the textural information in the image. It assumes that this textural information

is contained in the overall average spatial relation between grey tones in the image. A grey level co-occurrence matrix is determined which counts the occurrences of neighbouring pixels to the reference pixels given a certain distance. For example let $Q = (dx, dy) = (0, 1)$ be the distance, meaning the pixel directly to the right is the neighbouring pixel. For all the pixels of a given reference intensity, the amount occurrences an intensity in the neighbouring pixel is then captured in the grey level co-occurrence matrix $P(i, j)$ as displayed in 3-6.

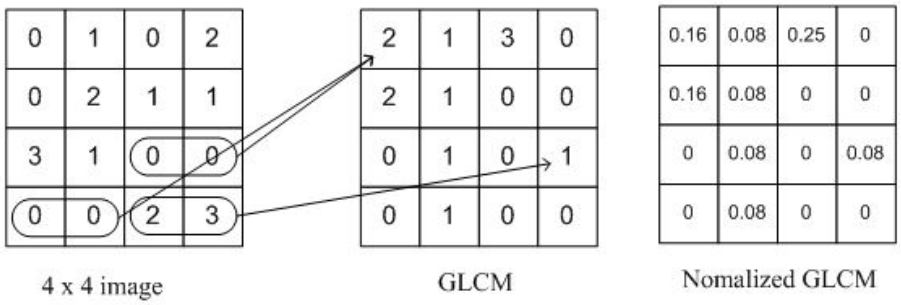


Figure 3-6: Example of determination of normalised GLCM from intensity image [12]

This matrix is then used to determine the various Haralick descriptors. The most commonly used are maximum probability, correlation, contrast, energy, homogeneity and entropy as the other descriptors are all correlated to these. They are described by

$$MP = \max_{i,j} P'(i, j) \tag{3-9}$$

$$CR = \Sigma_i \Sigma_j \frac{(i - \mu_x)(j - \mu_y)}{\sigma_x \sigma_y} \tag{3-10}$$

$$CN = \Sigma_i \Sigma_j (i - j)^2 P(i, j) \tag{3-11}$$

$$EN = \Sigma_i \Sigma_j P^2(i, j) \tag{3-12}$$

$$HO = \Sigma_i \Sigma_j \frac{P(i, j)}{1 + |i - j|} \tag{3-13}$$

$$ET = -\Sigma_i \Sigma_j P(i, j) log_2 P(i, j) \tag{3-14}$$

where

$$\mu_x = \Sigma_i \Sigma_j i P(i, j) \tag{3-15}$$

$$\mu_y = \Sigma_i \Sigma_j j P(i, j) \tag{3-16}$$

$$\sigma_x = \sqrt{\Sigma_i \Sigma_j (i - \mu_x)^2 P(i, j)} \tag{3-17}$$

$$\sigma_y = \sqrt{\Sigma_i \Sigma_j (i - \mu_x)^2 P(i, j)} \tag{3-18}$$

3-9-3 Classifier

As stated before, since the dataset is limited in size, a classifier is needed that has good performance with relatively many features compared to the amount of samples to train on.

Two classifiers are particularly well suited for this. The support vector machine and the random forest classifier.

Support Vector Machine

The first classifiers is the support vector machine [13]. In its simplest form it just finds the linear decision boundary that perfectly separates the classes with the largest margin as displayed in 3-7

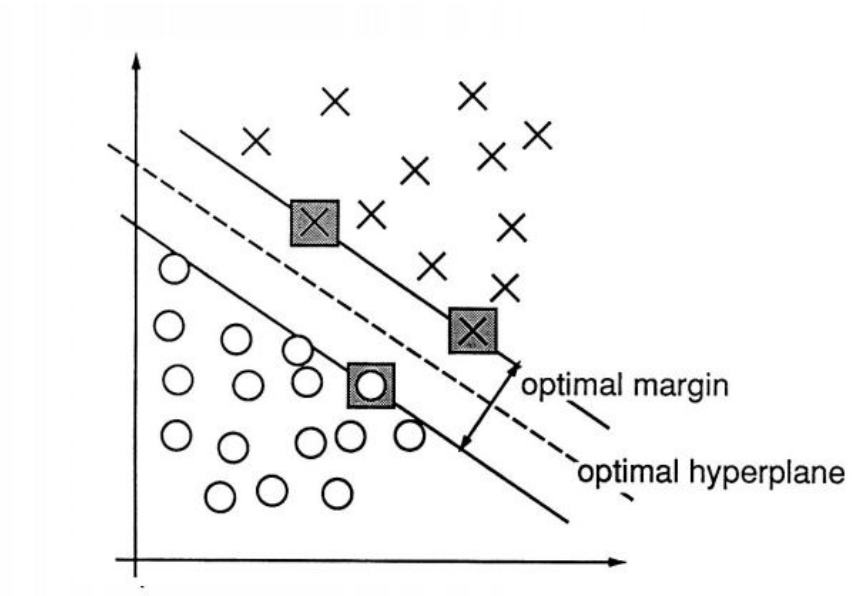


Figure 3-7: linear decision surface in 2 dimensional space with largest margin [13]

The respective optimization problem is

$$\begin{aligned} & \min \frac{1}{2}\|\boldsymbol{w}^2\| \\ \text{s.t.} \quad & w^T x_i + b \geq +1, \quad \text{for } y_i = +1 \\ \text{s.t.} \quad & w^T x_i + b \leq -1, \quad \text{for } y_i = -1 \end{aligned}$$

Introducing slack variables and solving the Lagrangian leads to

$$\begin{aligned} & \max_{\alpha} \Sigma_{i=1}^{N} \alpha_i - \frac{1}{2} \Sigma_{i,j}^{N} y_i y_j \alpha_i \alpha_j \boldsymbol{x}_i^T \boldsymbol{x}_j \\ \text{s.t.} \quad & \alpha_i \geq 0 \quad \forall i \\ & \Sigma_{i=1}^{N} \alpha_i y_i = 0 \quad \forall i \end{aligned}$$

where α_i is the lagrange multiplyer. Finally, using of a non-linear mapping into a high dimensional feature space and finding the linear decision surface that separates the classes in this higher dimensional space, makes nonlinear boundaries in the original feature space possible. However, working in high dimensional spaces is computationally very heavy. In order to prevent this a kernel is used, which prevents the need for explicit mapping in the higher order space. Given that the kernel $k(x_i, x_j)$ is equal to the inner product $< \phi(\boldsymbol{x_i}), \phi(x_j) >$, the kernel can be used in the optimization problem instead. This leads to the following optimization problem.

$$\begin{aligned} & \max_{\alpha} \Sigma_{i=1}^{N} \alpha_i - \frac{1}{2} \Sigma_{i,j}^{N} y_i y_j \alpha_i \alpha_j K(\boldsymbol{x}_i^T \boldsymbol{x}_j) \\ \text{s.t.} \quad & \alpha_i \geq 0 \quad \forall i \\ & \Sigma_{i=1}^{N} \alpha_i y_i = 0 \quad \forall i \end{aligned}$$

and the class of a new sample given by

$$f(z) = \Sigma_{i=1}^{N} \alpha_i y_i K(\boldsymbol{x}_i, \boldsymbol{z}) + b$$

where $\boldsymbol{z}$ is the feature vector of a new sample. The support vector machine has been shown to be a very powerful classifier, able to generalise very well. It is thus very suitable for situations where the feature space is relatively large for the amount of samples to train on.

Random Forrest

The other classifier to be discussed is the random forest classifier [44]. It is based on the simple decision tree classifier. In this classifier the data is sequentially split along a single feature as shown in 3-8. The set is split at the value which gives the lowest inter-class similarity and highest intra-class similarity. It will continue to do so until a stopping criterion is met. However, these decision tree classifiers are very simple and depending on the stopping criterion often not complex enough. Random forest tries to overcome this problem by training many slightly different decision trees. These different trees are made using boosting and bagging. Boosting is the process of altering the tree by for example changing the tree depth or altering the features available to the tree, for example having the tree train on only 5 randomly selected features out of 10. Bagging in the process of training multiple trees on random subsets of the data. These subsets are with replacements, meaning a single sample can be present multiple times in this new subset. Having created this large amount of different decision tree classifier, the random forest classifier works by majority rule. The decision tree classifier generally has a low bias and both boosting and bagging can greatly reduce the variance, leading to a very powerful classifier. Random forest classifiers have even been shown to be able to outperform support vector machines at large numbers of trees, though they are generally computationally heavier at these amounts.

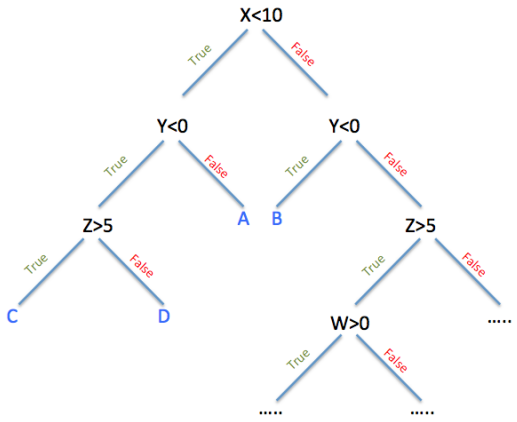


Figure 3-8: Example of a decision tree [14]

3-10 The Microscope

The microscope used to make the hypserspectral images is very similar to a conventional bright-field microscope. As displayed in 3-9, it has an illumination path consisting of a light source, collector lens, field aperture, field lens, condenser aperture and a condenser lens, of which the latter is adjustable to be able to align it so that koehler illumination can be achieved. The sample substage consist of a holder in which the glass slides can be fixed in place, which itself is xy-tunable to be able to examine different parts of the sample. The capturing stage consists of a objective lens which is y-tunable to allow for focusing on the sample, field lens and camera. Three of these parts are somewhat different from what you would find on a regular bright-field microscope and will be treated more in depth. These are the light source, the objective lens and the camera.

3-10-1 Light Source

The first and most important difference between a conventional bright-field microscope and this hyperspectral microscope is the light source. It consists of three part. The first is a halogen lamp which emits light in a broad spectrum. This halogen lamp is followed by a monochromator, which is used to select a much more narrow band of wavelengths from this wider spectrum. It does so through dispersion, as can be seen in 3-10. The width of the slid determines the width of the resulting waveband. The narrower the slid, the narrower the waveband and thus the higher the spectral resolution. However, it also greatly reduces the brightness of the illumination and thus reduces both the signal to noise ratio and how much of the available spectrum from the halogen lamp can be used. A slid of 300 micro meter was found to be a good trade-off, being able to use most of the provided spectrum with a decent signal to noise ratio, yet retaining a high spectral signature, wavebands of 4 nano meter wide having minimal overlap.

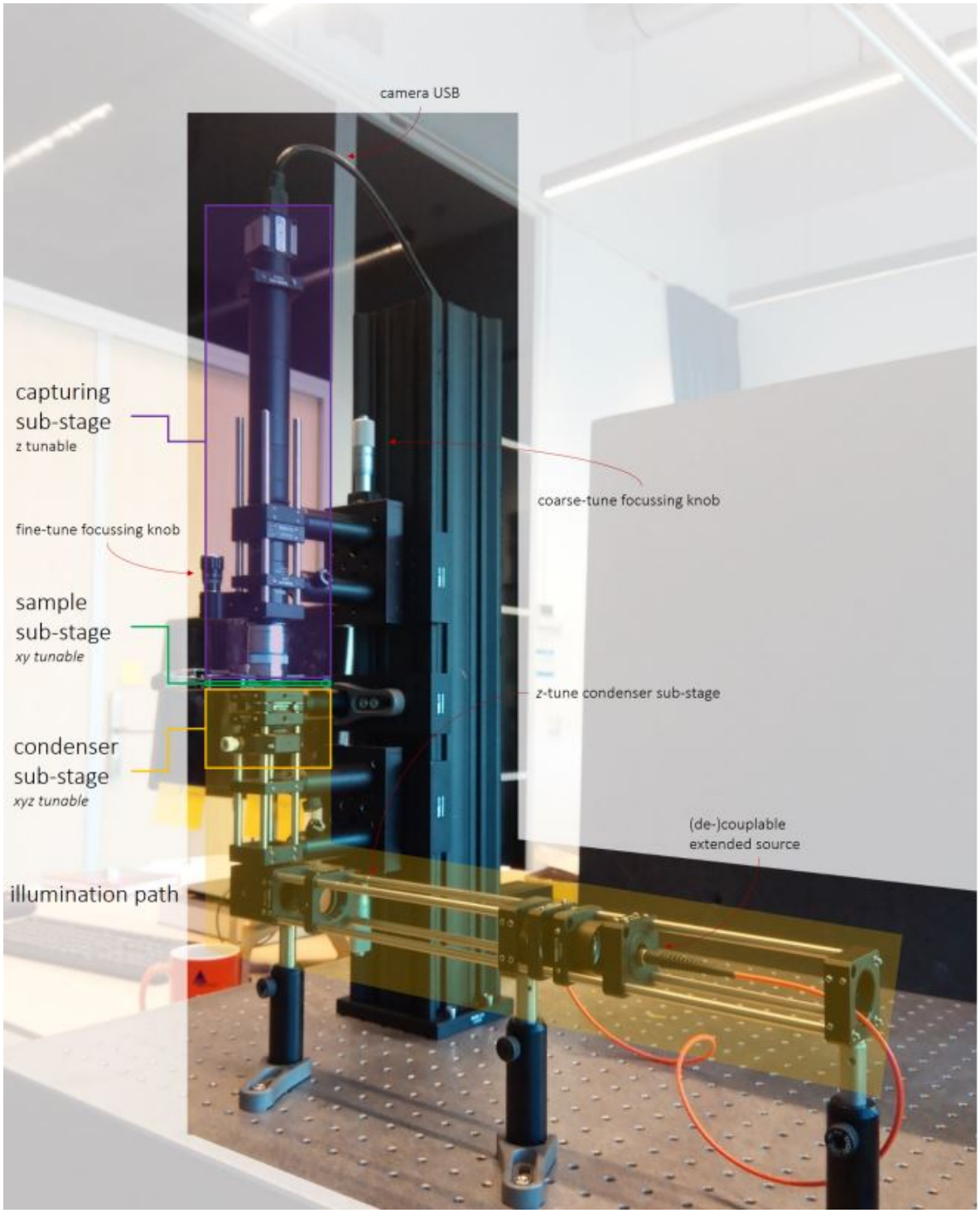


Figure 3-9: The hyperspectral microscope used in the experiments[1]

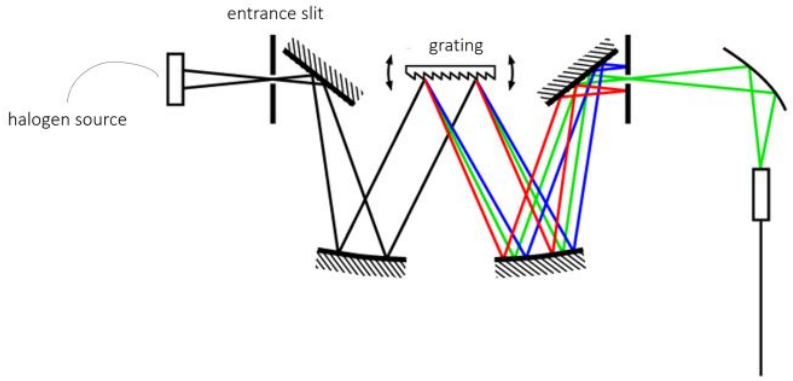
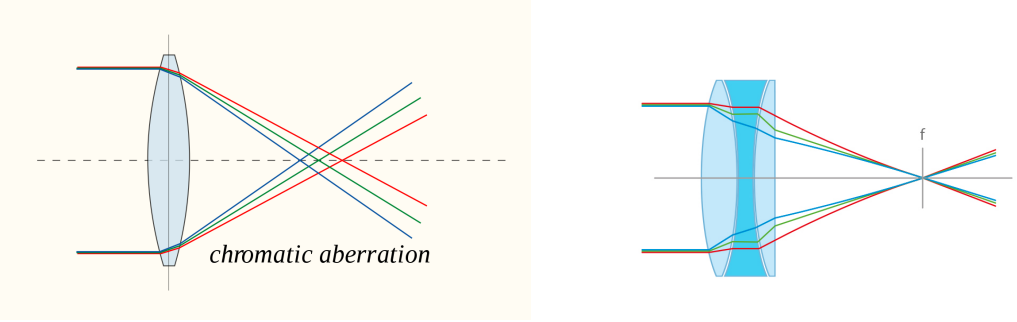


Figure 3-10: The schematic of a monochromator. The entrance slid determines how narrow the exiting waveband is and turning the grating changes the outgoing wavelength

Finally, the light is captured by a fiber optics cable which lead to the collector lens of the microscope.

3-10-2 Objective Lens

Next is the objective lens for which a apo-chromatic lens has been chosen. This is due to what is called chromatic aberration, a phenomenon which causes the sample to be in focus when illuminated with one wavelength, yet out of focus when illuminated with another. An apo-chromatic lens has been developed in such a way as to minimise this effect in the visible spectrum, as can be seen in 3-11.



(a) (b)

Figure 3-11: Chromatic aberration causes different wavelengths to have different focal points (a). In apochromatic lenses three wavelengths of interest have the same focal point with minimal defocus in between (b). [15]

3-10-3 Camera

Lastly, instead of an RGB sensor which captures an intensity value for red, green and blue simultaneously, a monochromatic camera is used as only a very small part of the spectrum is used for illumination at a time. This way only a single intensity image is created at a time and by changing the wavelength of the light source each time before the next image is taken the hyperspectral image of the sample is captured.

Chapter 4

Results

This chapter outline how the methods described in the previous chapter are applied in this research. More specifically, how they are applied for the purpose of parasite detecion in stained Malaria infected blood samples, unstained Malaria infected blood samples and finally Schistosoma infected urine samples. First the methodology of how these methods are used and how they are evaluated is explained. After this a more detailed description of each experiment is given and the results are presented. The largest part of the chapter is dedicated to the analysis of the endmember extraction methods. These do not have a groundtruth to compare the results to and their performance analysis thus poses the biggest challenge. The smaller part is dedicated to the analysis of a hypotetical multispectral setup which uses the knowledge derived in the endmember extraction. This setup is compared to traditional RGB imaging to evaluate the potential multispectral imaging has in the field of parasite detection.

4-1 Methodology

This section will start by mentioning the preprocessing used on the data and the parameters used in the various methods. It will then explain the methodology used to determine the performance of the various endmember extraction methods and classification schemes. More detailed descriptions of the experiments are given throughout the chapter at the start of the corresponding sections. It will conclude with a quick overview of the abbreviations used in this chapter.

4-1-1 Technical Details

Firstly, the parameters and preprocessing used in the experiments will be briefly mentioned. The data is treated for the sensor disturbance as was explained in 3-2-1. The illumination correction is applied to the endmembers in order to accurately represent them. Furthermore, it was found that hot and dead pixels greatly affect the performance of the endmember extraction methods. Therefore, a 3 x 3 x 3 median filter is applied to the hypercube prior

to any of the denoising steps. This has a negative effect on the ability of the endmember extraction to be able to pick up on subpixel substances, but it was found to improve both the separation of parasites, RBC's and background as well as improve classification performance of the reference spectrum assisted classification. Of course, this median filter already has a denoising characteristic. Therefore, the endmembers extraction methods are also tested without any further denoising. When further denoising is applied the following parameters are used. The 3d gaussian filter was chosen to be 3 x 3 x 3 and the savitzky golay filter was given a window size of 11 and a polynomial degree of 4. The dimensionallity reduction methods did not require additional parameters to be chosen other than the required dimensionallity. The number of skewers used by the PPI algorithm was chosen to be the industry standard of 10000 as this gave it a similar computational load as NFINDR, which did not require additional parameters. Statistics based and Sisal also did not require additional parameters to be chosen.

4-1-2 Hyperspectral - RGB Comparison

The first experiment that has been conducted was that of comparing the classification results using the hyperspectral data directly as compared to using RGB images. To make the comparison as fair as possible the RGB data is derived from the hyperspectral data by averaging over the corresponding wavelengths. Since the light source used in the microscope was found to not be bright enough in the lower wavelengths of the blue spectrum the choice was made to only use red and green. Using a spectral resolution of 4 nano meter this resulted in 50 images ranging the wavelengths from 500 to 700 nano meter for the hyperspectral data and 2 images, 500-600 and 600-700, for the RGB data. Finally, both sets are reduced to 1 image (each) using PCA and the same set of features, namely mean, variance, dissimilarity and contrast, is derived. Using the same classifier the results will give valuable insight into the potential of hyperspectral imaging in this application.

4-1-3 Spectral Signature Validation

Next, the endmember extraction methods are validated. The endmember extraction is an unsupervised process and there is no groundtruth to compare either the resulting spectra or corresponding abundance maps to. To determine their performance four methods are used. In the first step the locations of the endmembers in the image are inspected. For the pure pixel methods the pure pixels are used and for the other the most similar pixels are determined using the SAM metric on the median filtered image. Secondly, the spectra themselves are inspected to see if they demonstrate unlikely or noise like behaviour. Thirdly, FCLS is used to create the corresponding abundance maps. Good performance in each of these tests are indicated by proper separation of the parasites (or parasite eggs), cells and background. These tests are conducted for the stained Malaria sample, unstained malaria sample and Schistosomiasis sample. Finally, in case of the Malaria sample, the spectral signatures are used to create detection maps from which features are then derived. The performance of the classification using these features will then be a good indicator of the quality of the spectral signatures. By extension, they will also be a good indicator of the corresponding endmember extraction method. These performances are furthermore compared to the red-green and hyperspectral image classifiers mentioned previously.

4-1-4 Multispectral - RGB Comparison

Finally, the spectral signatures corresponding to the most successsful endmember extraction method are used to determine the wavelengths which give the most discriminative power. A hypothetical multispectral setup imaging only at the two most discriminative wavelengths is compared to red-green imaging. To make sure both have a similar signal-to-noise ratio both are derived from the same hyperspectral data and for both a single image is constructed by averaging five underlying images in the hyperspectral data. For the multispectral setup these are the image at the wavelength and the four neighbouring it. For the red-green images these are five wavelengths equidistant from each other in the relevant waveband are used. Similar to before these sets are reduced in dimensionallity using PCA. The same set of features are used for classification, namely mean, variance, dissimilarity and contrast.

4-1-5 Abbreviations

To refer to the various combinations of denoising, virtual dimesnionallity, dimensionallity reduction and endmember extraction methods described in the previous section some abbreviation will be used, as displayed in table 4-1-5. So for example, the combination of savitzky golay filting, a virtual dimensionallity of 4, pca dimensionallity reduction and PPI endmember extraction will be denoted by S4PP. One notable exception is when Sisal is used as the dimensionallity reduction is part of the endmember extraction process. A combination using Sisal, no denoising and a VD of 3 would be N3S. Why HySime is grouped with the dimensionallity reduction instead of the virtual dimensionallity is explained in 4-3-1.

Noise reduction	virtual dimensionallity	dimesnionallity reduction	endmenber extraction
N (None)	3	P (PCA)	P (PPI)
G (3d Gaussian)	4	I (ICA)	N (NFINDR)
S (Savitsky Golay)	5	H (Hysime projection)	B (Statistics based)
	6		S (Sisal)

4-2 Hyperspectral - RGB Comparison

The first experiment that has been conducted was to compare hyperspectral imaging and RGB imaging directly in the the application it is to be used, namely classification. The aim is to determine on individual cells whether they are infected or not. This is done by comparing two classifiers, one trained on features derived from the hyperspectral data and the other trained on the same features but derived from the RGB data. However, the used lightsource does not extend into the blue spectrum with enough brightness so only red and green are used. Since the spectral range is limeted for both the comparison is still informative and will likely extend to the situation where the blue waveband is included. Both the hyperspectral data and red-green images are derived from the same set of images ranging from 500 to 700 nano meter. The RBC's are segmented from these images and labelled as being infected or not. The dimensionallity of both datasets is reduced to 1 image by the PCA transform as

the lower feature count greatly improved the classification performance. The features that are derived from these images are the mean, variance, contrast and dissimilarity as analysis of both classifiers showed these features to be the most important. The classifiers are the SVM with a regularization parameter of 1000 and a balanced class weight and random forest with 100 individual trees and also a balanced class weight. They are trained and tested using leave-one-out crossvalidation. The results are given in table 4-1.

	sens	spec	s
PCA-SVM	0.696	0.888	0.861
RG-SVM	0.867	0.867	0.849
PCA-Forest	0.478	0.972	0.903
RG-Forest	0.565	0.979	0.922

Table 4-1: The sensitivity, specificity and success rate of the SVM and Random Forest classifiers using the red-green images or first two principal components.

Interestingly, the red-green simulated images provided a better basis for classification than the full hyperspectral data as it resulted in a significantly higher sensitivity in both classifiers. An unexpected result since the RGB image is derived from the hyperspectral data and holds much less information. This implies that the principal component analysis generalises to much in the process of turning the 75 wavebands into 1 image. However, due to the relatively small samplesize, adding more features or using supervised dimensionallity reduction such as LDA only decreases the performance, likely due to overtraining. In order to utilize the extra information in the hyperspectral data thus requires either a higher samplesize or some other way to reduce the dimensionallity. Since the first option is not viable at this time due to the slow data acquisition the second option is explored further. Using a priori knowledge derived in the endmember extraction the classifier using the hyperspectral data will be shown to outperform the classifier using the RGB data in section 4-4.

4-3 Spectral Signature Validation

The next step is to determine whether the endmember extraction methods combined with the various methods of virtual dimensionallity estimation, denoising and dimensionallity reduction are able to accurately estimate the spectral signature of the parasitic subtance in the samples. This is tested on firstly the unstained Malaria parasites, secondly the Schistosoma eggs and thirdly the unstained Malaria parasites. The output of these methods is a set of spectra for which the groundtruth is unknown. This makes it significantly harder to determine the performance. Luckily, there are also multiple things that are known, the first being the endmember locations. Of these at least one should be located on a parasite, one on a RBC and one on the background. In case of the pure pixel methods the endmembers correspond to a pixel in the image. In case of statistics based and Sisal the Spectral Angle Mapper metric is used to determine the pixels in the image which are most similar to the endmembers. Next the spectra themselves can be inspected. It is known that one endmember should have high light transmittance values for all wavelengths, namely the one that belongs to the background where there are no cells or parasites. Furthermore, any sudden jumps or noise like behaviour serves as an indicator that it is not an actual spectral signature of a substance in the sample.

Thirdly, the FCLS regression method is used to create abundance maps corresponding to each of the endmembers. Here, clear low noise abundance maps in which the parasites, RBC's and background are separated well from each other are an indicator of good performance. These three validation methods are also applied to the Schistosoma egg infested Schistosoma sample and the unstained Malaria infected blood sample. Furthermore, each sample is imaged using a 20x apochormatic magnification objective lens a spectral range of 500nm to 750 nm was chosen. Below this the lightsource was not be strong enough to provide a good signal-to-noise ratio and above this large chromatic aberration was found to occur.

4-3-1 Virtual Dimensionallity Estimation and Dimensionallity Reduction

The first step in the spectral signature estimation process is the estimation of the virtual dimensionallity. For this the HySime method was chosen. However, the method was found to consistently overestimate the VD no matter the preprocessing used. It often found the VD to be 22 or higher. However, when the data is projected onto this subspace. All the dimensions past the fifth were found to mostly contain noise, as is displayed in 4-1.

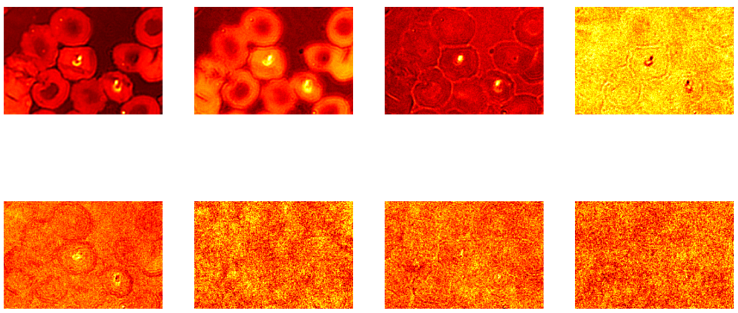


Figure 4-1: The first 8 dimensions of the HySime subspace

This was also the case when other dimensionallity reduction was used, as can be seen in 4-2 and 4-3.

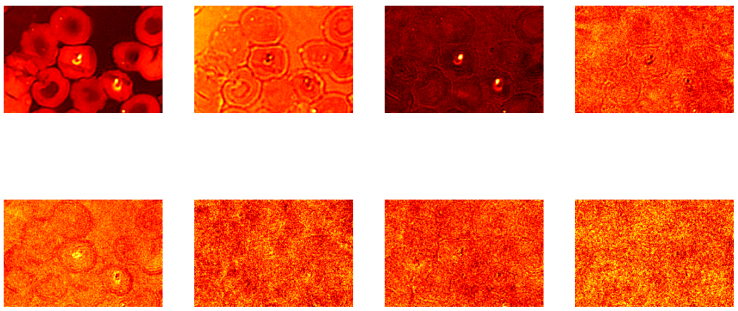


Figure 4-2: The first 8 components of the PCA transform

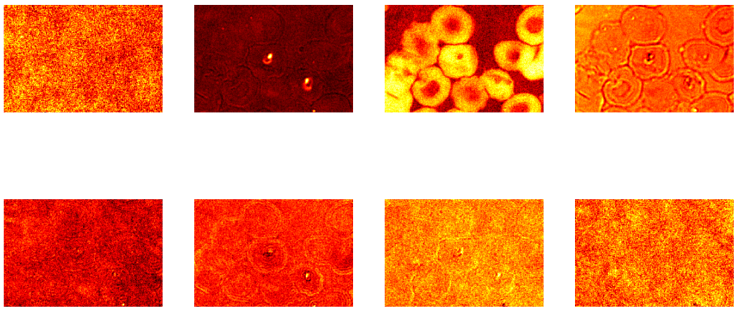


Figure 4-3: The first 8 components of the ICA transform

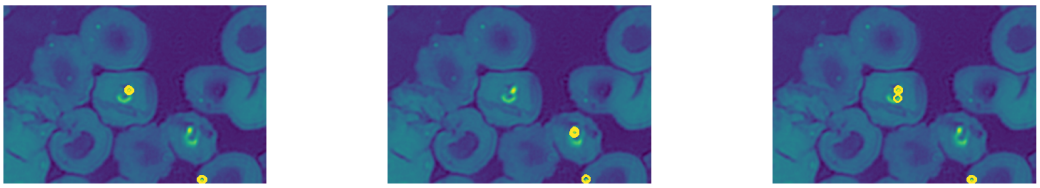
Therefore, the choice was made to test each method using a VD of 3 to 6 and compare them instead of relying on the HySime method for VD estimation. However, the HySime method does provide an additional way of dimensionallity reduction and is still used for that purpose in this research.

4-3-2 The Stained Malaria Infected Blood Sample

The first sample is the giemsa stained Malaria infected thin smear blood sample and should provide the easiest task as these parasites are already easily visible in regular RGB images. A part of the full image which contains some clearly visible parasites is cut. Note that the amount of pixels is still much higher than what is generally used in many other hyperspectral imaging applications such as remote sensing.

PPI

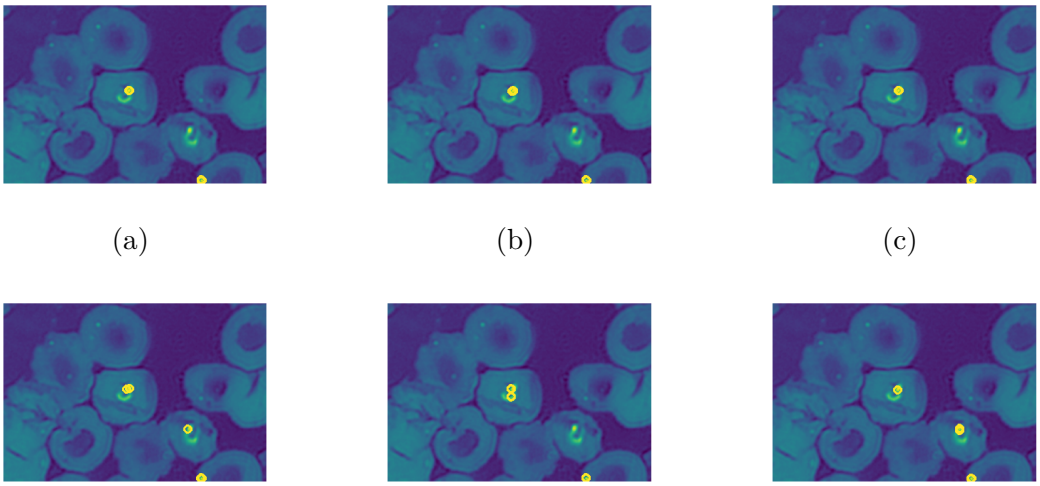
Due to how the PPI method works, the choice of a higher or lower VD does not impact the first endmembers that are found. However, since the skewers are generated randomly, each run could produce different results. Running several combinations multiple times these differences were found to be quite severe, as displayed in 4-4 where three runs of the same method produce significantly different results.



(a) (b) (c)

Figure 4-4: Three runs of the N6IP method resulting in different endmember locations highlighted by the yellow circles

Looking at general trends in behaviour, figure 4-5 a to f show that the pixels corresponding with the endmembers are consistently those belonging parasites. However, it often selects neighbouring pixels of the same parasite as separate endmembers. Furthermore, there seem to be no pixels that belong to the RBC's or other substances. Since a set of distinct spectral signatures containing a singular spectral signature corresponding to the parasite is sought after, the PPI method seems to be lacking in this regard.



(a) (b) (c)

(d) (e) (f)

Figure 4-5: The locations of the pure pixels for the N6PP (a), G6PP (b), S6PP (c), N6IP (d), G6IP (e) and S6IP (f) methods on the Malaria sample highlighted by the yellow circles

As the method was found to select neighbouring pixels as endmembers it is no surprise that the spectral signatures are similar also, as displayed in 4-6 a to d, which display the spectral signatures resulting from the PPI method when coupled with Savitsky Golay/ 3d gaussian filtering and PCA /ICA. In each example, there are only two distinctly different endmembers, the other endmembers being slight variations, neither meeting the requirements of a background spectral signature. This is the case for most combinations using PPI, with the occasional exception, but these exceptions are caused by the randomness of the PPI method rather than the specific combination of denoising and/or dimensionallity reduction method.

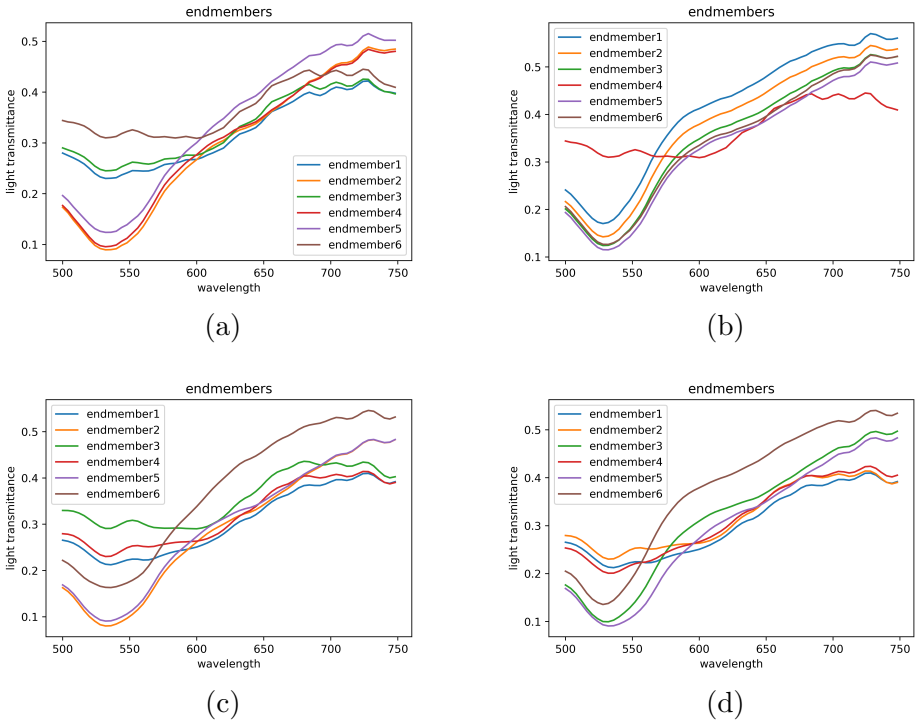


Figure 4-6: The spectral signatures of the endmembers found by the for the N6PP (a), G6IP (b), S6PP (c) and S6IP (d) methods on the Malaria sample highlighted by the yellow circles

However, when the PPI method seems to work well, the corresponding abundance maps show the parasite being separated from the rest of the image using only 3 endmembers, the best result being displayed in 4-7. However, even in this case the PPI method is not able to separate the RBC's and the background from each other. Nonetheless, the endmember that corresponds to the abundance map that most clearly shows the parasite could still be a good estimation of the spectral signature and the PPI method should not be discarded just yet. It does however make it very hard to determine at which wavelengths the signature is most different from those the other substances as these have not been accurately estimated.

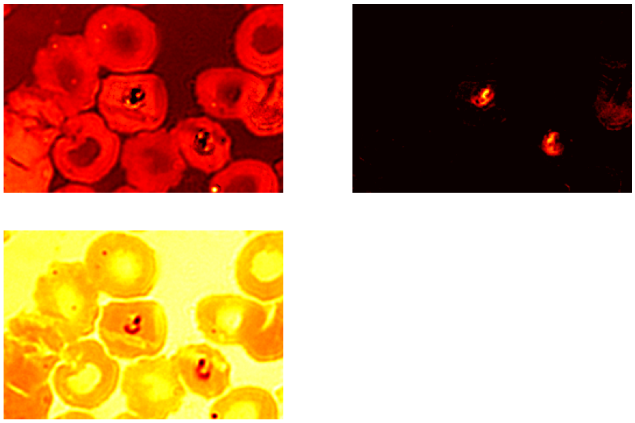
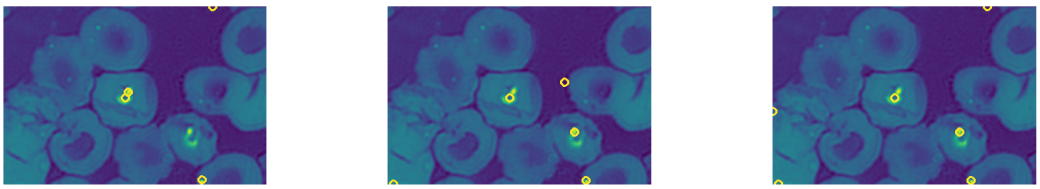


Figure 4-7: The abundance maps derived in the S3PP method

NFINDR

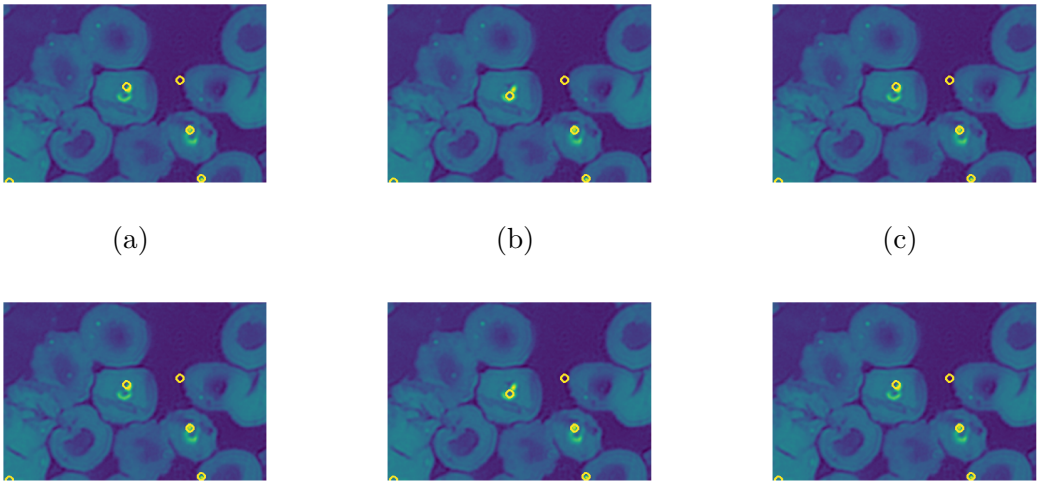
For NFINDR, unlike as was the case with PPI, multiple runs of the same method resulted in the same outcome. What did seem to have a large influence is the given virtual dimensionallity, as is displayed in 4-8. The locations of endmembers are very different when tasked with finding 4 endmembers compared to finding 6. However, looking at the corresponding spectral signatures, these remain very similar across different dimensionallities as can be seen in **??**.



(a) (b) (c)

Figure 4-8: Three runs of the G_PN method using different VD's resulting in different endmember locations

Eventually, a VD of 5 was chosen for the stained Malaria sample as this resulted the highest quality abundance maps. The NFINDR method clearly do not suffer from the same problems as the PPI method. Figures 4-9 a to f show no endmember pixels being neighbouring, though there are still two endmembers that seem to correspond to the parasites. It furthermore shows that it finds at least one endmember on a parasite, one on a RBC and one on the background, which is in line with the expectations.

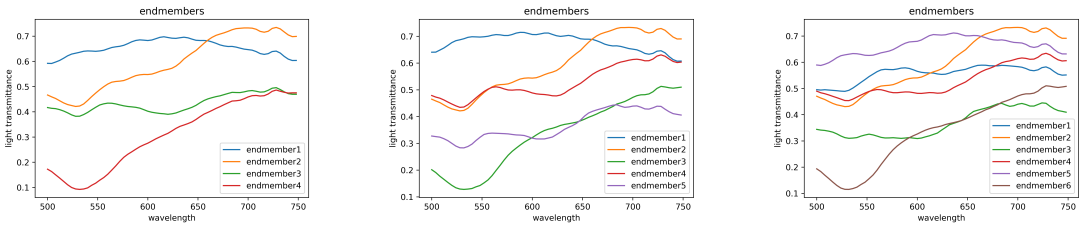


(a) (b) (c)

(d) (e) (f)

Figure 4-9: The locations of the pure pixels for the N5PN (a), G5PN (b), S5PN (c), N5IN (d), G5IN (e) and S5IN (f) methods on the Malaria sample highlighted by the yellow circles

Furthermore, the resulting spectral signatures are much more distinct from each other than those of resulting from PPI as can be seen in **??**. Each set contains one endmember which lets through all wavelengths equally, corresponding to parts of the sample where no RBC of parasite is present, thus being our sought after background pixel.



(a) (b) (c)

Figure 4-10: The spectral signatures of the endmembers found by the G4PN (a), G5PN (b) and G6PN (c) method

The corresponding abundance maps resulting from the NFINDR endmember extraction are are shown to be high detail and low noise, the best images being derived using the savitsky golay filtering, though the differences are small. However, the results vary significantly across different dimensionallities. Given a VD of 3 the method was not able to separate the parasite from the rest of the image while given a VD of 4 combined with HySime it was able to do so 4-11. A VD of 5 and using the PCA transform resulted in a clear separation of the parasites while containing much less noise than using HySime 4-12. Unlike PPI, NFINDR creates a clear separation between the parasite, the RBC's and the background. It is unclear what the remaining two abundance maps might correspond to, but it is clear that the higher VD has a beneficial effect on the results as a whole.

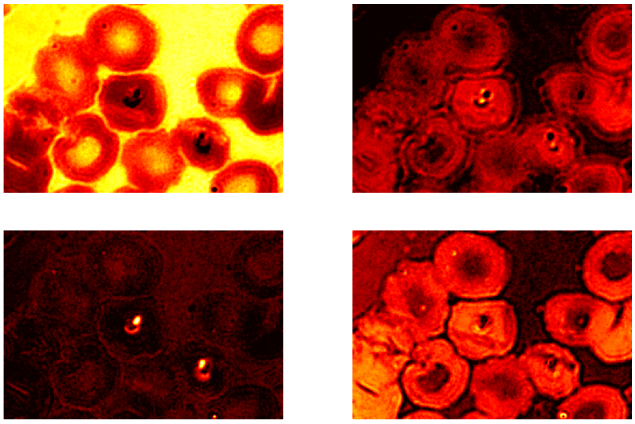


Figure 4-11: The endmembers derived in the S4HN method with negative light transmittance values

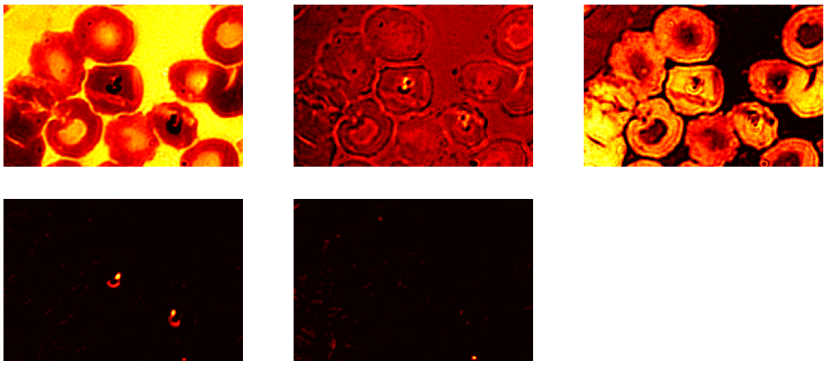
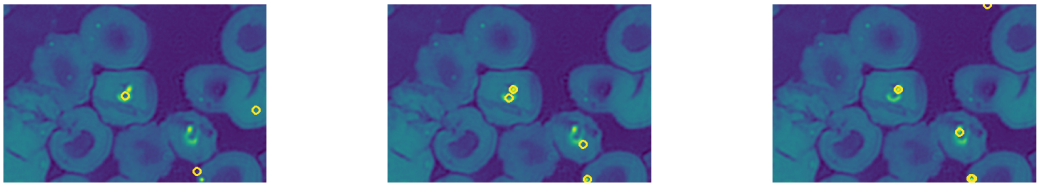


Figure 4-12: The endmembers derived in the S5PN method with negative light transmittance values

Statistics Based

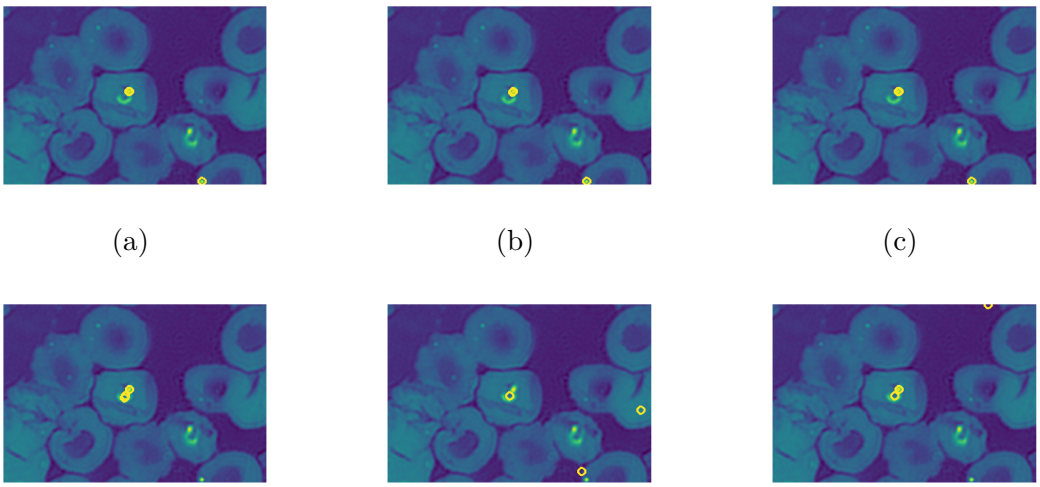
The statistics based method does not select pixels to be endmembers like the pure pixel methods. Therefore, the spectral angle mapper metric is used to determine the pixels which are most similar to the endmembers. Similar to NFINDR, the results are shown to be very dependant on the given VD, as can be seen in 4-13. However, it is hard to argue what VD works best reasoning from the pixel locations as most results are plausible. Given the ICA transform is used, there is generally a pixel corresponding to a parasite and the other pixels are generally distinct from each other.



(a) (b) (c)

Figure 4-13: Three runs of the G_IS method using different VD's resulting in different endmember locations

Another important factor is the dimensionallity reduction. Next to finding neighbouring pixels, due to finding the pixels using SAM it is also possible to find the exact same pixel for multiple endmembers. This is shown to often be the case when PCA is used, as can be seen in 4-14. The method works much better when coupled with the ICA transform, finding many multiple distinct pixels. A VD of 3 was chosen as this results in the best abundance maps. Finally, the choice of denoising method also has a significant effect on the results. When 3d gaussian filtering is used, the method is often able to find pixels on the parasite, the RBC's and the background, compared with 3 pixels on the parasite when no denoising is applied.

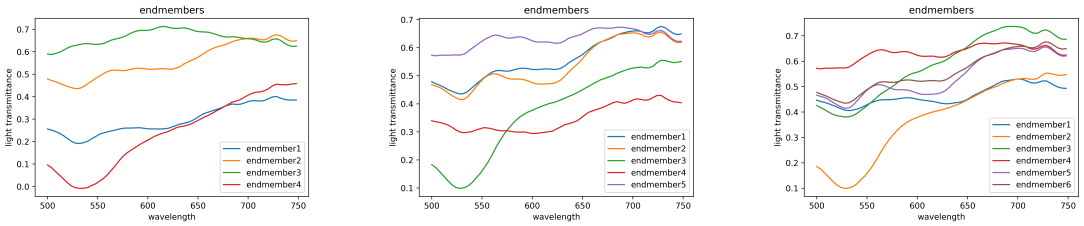


(a) (b) (c)

(d) (e) (f)

Figure 4-14: The locations of the pure pixels for the N3PB (a), G3PB (b), S3PB (c), N3IB (d), G3IB (e) and S3IB (f) methods on the Malaria sample highlighted by the yellow circles

When the ICA transform is used, though the endmember locations differed, the spectral signatures are shown to remain quite similar across different virtual dimensionallities, as can be seen in 4-31. Furthermore, using the ICA transform, each set of endmembers is shown to have a spectral signature with high values at all wavelengths, which can thus be interpreted as the background pixel.



(a) (b) (c)

Figure 4-15: The spectral signatures of the endmembers found by the G4PN (a), G5PN (b) and G6PN (c) method

The same story holds for the abundance maps. The method combined with PCA consistently produces low information abundance maps with one almost completely white except for the location of the parasite, which is displayed in another. Using a VD of 3 and the ICA transform the method is able to derive abundance maps in which the parasite was separated from the rest of the image, especially when combined with the Savitsky Golay filtering. However, the RBC's and background are still not properly separated, as can be seen in 4-16. The inability to separate the RBC's and the background is likely due to the RBC's higher abundance and thus higher variance leading to the information on them mostly being lost in the ICA transform.

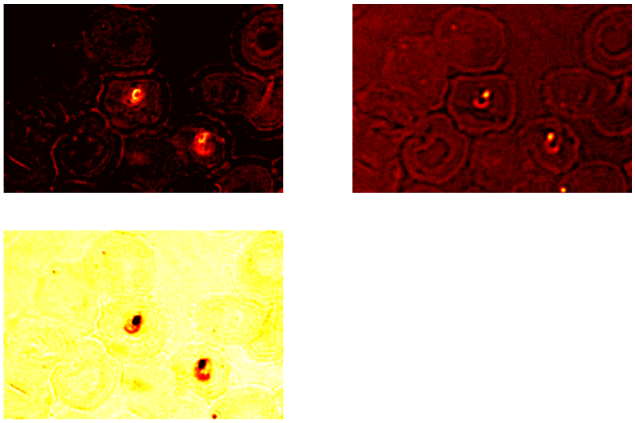
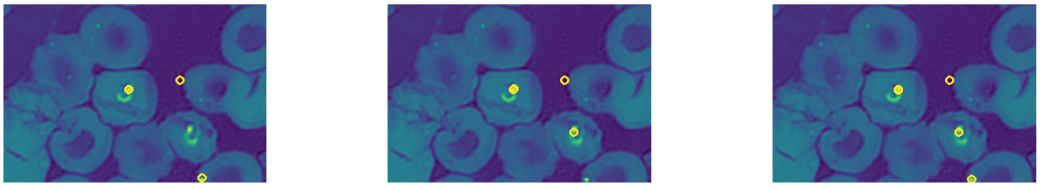


Figure 4-16: The endmembers derived in the S3IB method with negative light transmittance values

Sisal

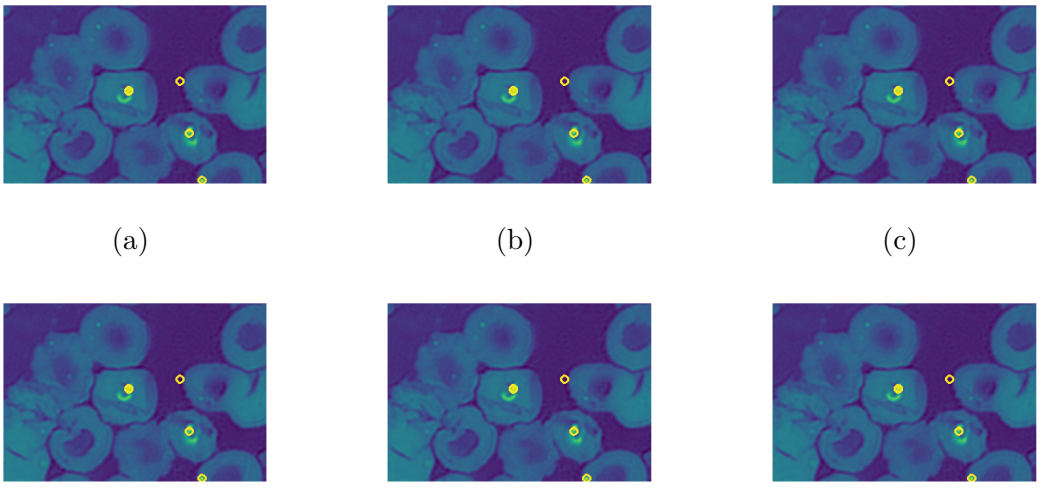
Similar to the statistics based method, the endmembers of the Sisal method are evaluated using the SAM metric. Again, the results can change depending on the given VD. This is shown to be the case in 4-17.



(a) (b) (c)

Figure 4-17: Three runs of the G_S method using different VD's resulting in different endmember locations

The results of the method with a VD of 5 are shown as these generally produce 4 non neighboring pixels, while a VD of 4 results in only 3 as well as producing the producing the clearest abundance maps. The results shown in 4-18 clearly shows that of the corresponding pixels two are located on the upper parasite, one on the edge of the lower parasite and one on the edge of a RBC. Furthermore, it is shown that the choice of denoising and dimensionallity reduction has little to no effect on the location of these pixels.

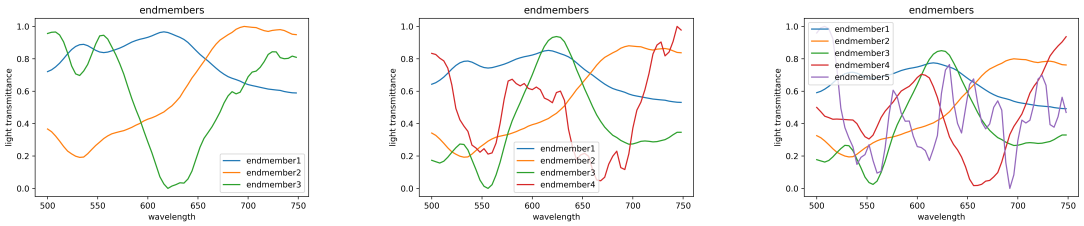


(a) (b) (c)

(d) (e) (f)

Figure 4-18: The locations of the pure pixels for the N5S (a), G5S (b), S5S (c), N5S (d), G5S (e) and S5S (f) methods on the Malaria sample highlighted by the yellow circles

Unlike other method, the endmember locations being very dependant on the given virtual dimensionallity is also reflected in the corresponding spectra, as can be seen in 4-19. The the difference between a VD of 3 and 4 is particularly notable as their third endmembers are the polar opposite of each other.



(a) (b) (c)

Figure 4-19: The spectral signatures of the endmembers found by the G3S (a), G4S (b) and G5S (c) method

The abundance maps resulting from the Sisal endmember extraction are in line with expectations given some endmembers were shown to have the same pixel as most similar using the spectral angle mapper metric. Using a VD of 4 or above, there are consistently 3 abundance maps which clearly show something, the remaining maps being mostly black, as is shown in 4-20. These 3 abundance maps clearly show one having a high abundance for the background, another having a high abundance for the RBC's as well as the parasite and the last showing the parasites, though slightly less bright. The parasites on this final abundance map are much brighter given a VD of 4 or higher compared to a VD of 3. This likely has to do with the third endmember flipping given a VD of 4 or above.

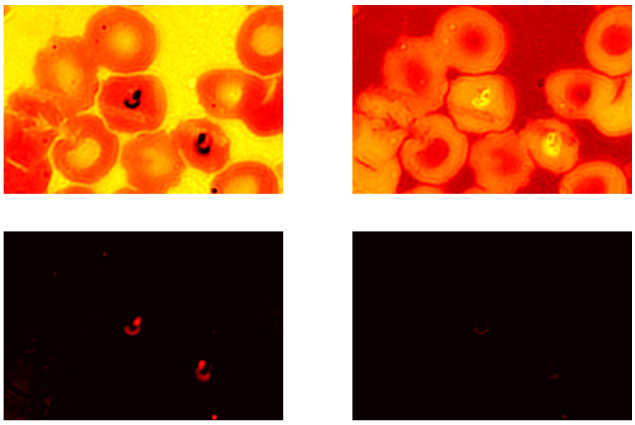


Figure 4-20: The endmembers derived in the G4S method with negative light transmittance values

4-3-3 Schistosoma Infected Urine Samples

The methods are tested in a similar manner on urine samples containing Schistosoma eggs. The eggs are relatively large compared to the Malaria parasites, the eggs of the species used for this experiment being visible using a 4x magnification objective lens. However, for these

experiments the sample has been imaged using a 20x magnification objective lens, resulting in an image only just large enough to contain the egg. The 20x lens is not very useful in a practical setting because it takes a long time to find the egg and thus does not make for an efficient diagnostics tool. However, it will likely make for a better approximation of the spectral signature, which can then in turn be used on the data corresponding to the 4x magnification objective lens. One final notable difference is that due to the eggs being thicker than the cells, only one of the 2 can be in focus at a time. Since the eggs are of interest the most, these will be set in focus.

PPI

The endmember locations of the PPI method are quite similar to when applied to the stained Malaria sample, as is shown in 4-21. The PPI method mostly selects pixels of the egg to be endmembers when coupled with PCA, while using ICA the method tends to also select pixels on the main white blood cell to the right to be endmembers. Using the ICA transform it furthermore has a lower tendency of selecting neighbouring pixels as endmembers. Whether it selects a background pixel is a bit harder to say since both the egg and the white blood cell are mostly transparent and might thus contain a pixel which could be a good representation of the background spectral signature. A VD of 4 was chosen as this resulted in one set of high quality abundance maps when combined with no denoising and ICA. Again, the method was found non-reliable as re-runs gave different results.

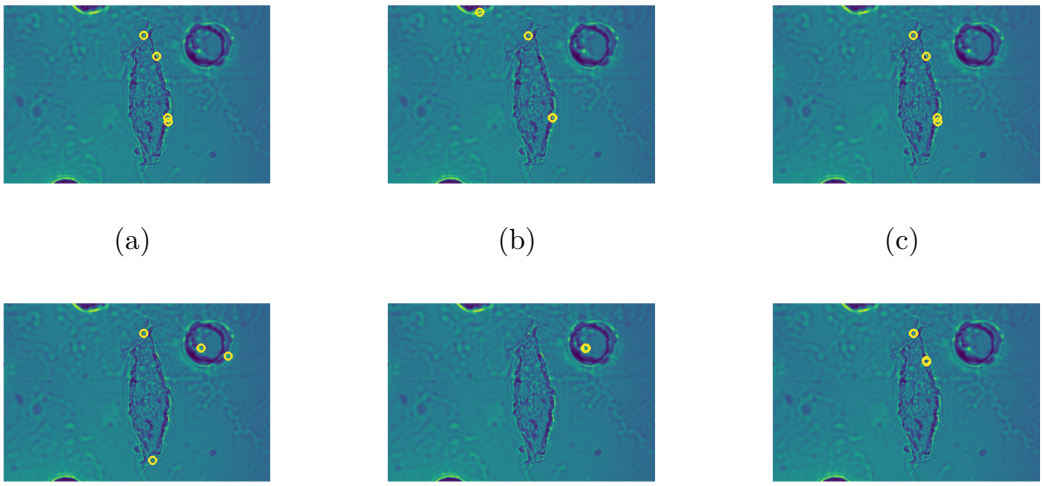
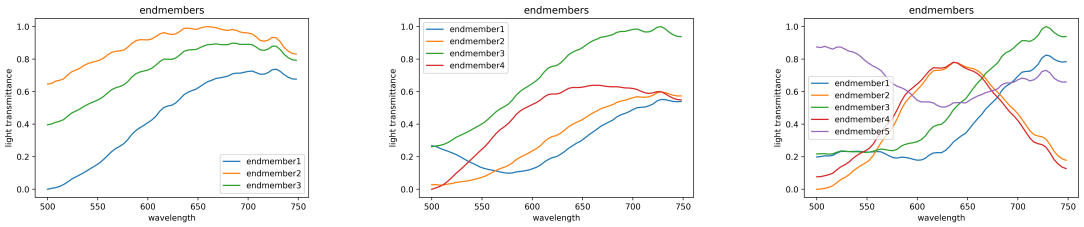


Figure 4-21: The locations of the pure pixels for the N4PP (a), G4PP (b), S4PP (c), N4IP (d), G4IP (e) and S4IP (f) methods on the Schistosoma sample highlighted by the yellow circles

The endmembers found by the PPI endmember extraction for the Schistosoma infected urine sample suffer from the same problems as those in the stained Malaria infected blood sample, though to a lesser extend. There is still a tendency to select multiple very similar endmembers,

but it is occasionally able to find 3 or even 4 distinct endmembers coupled with ICA, as was the case for the previously mentioned run, shown in 4-22.



(a) (b) (c)

Figure 4-22: The spectral signatures of the endmembers found by the G3IP (a), G4IP (b) and G5IP (c) method

This specific result also corresponds to a set of high quality abundance maps, as can be seen in 4-23. The egg is separated from the background very well, though it is not able to separate the egg from the white blood cell.

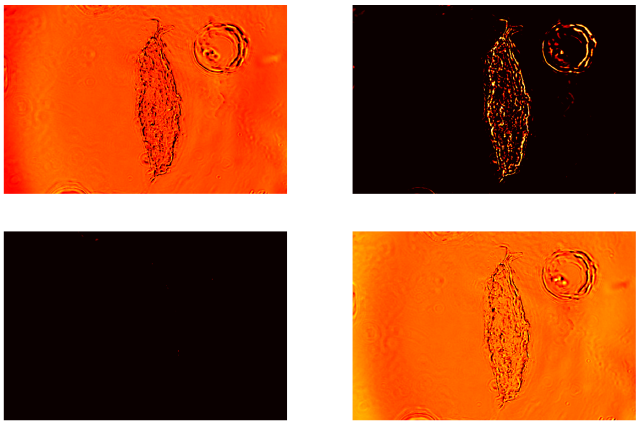
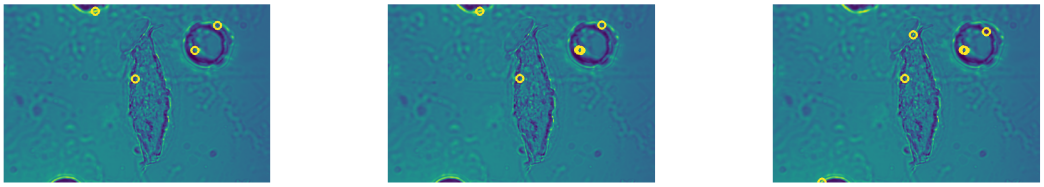


Figure 4-23: The endmembers derived in the S4IP method with negative light transmittance values

NFINDR

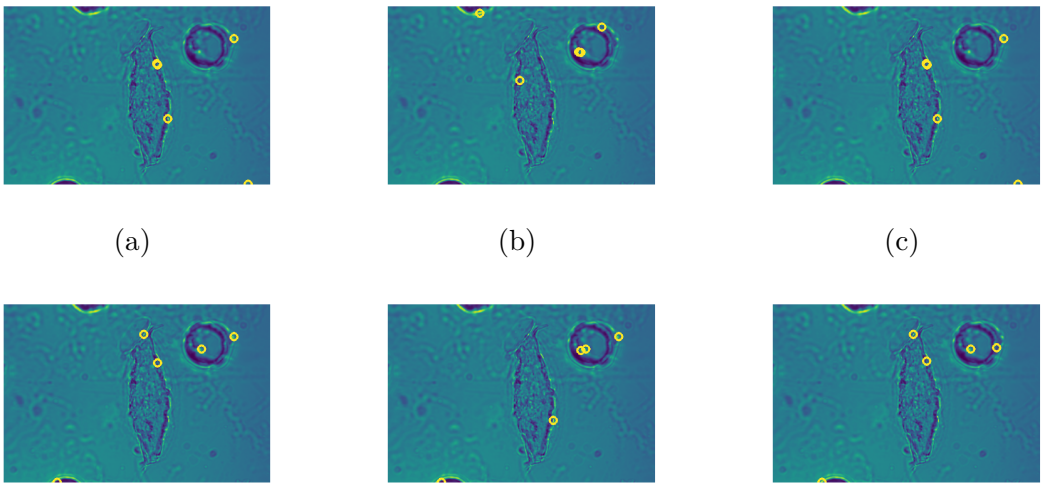
The NDINDR method is much more consistent and only rarely selects neighbouring pixels as endmembers. However, similarly to PPI, there is no clear background pixel being selected, though this could again be caused by the transparency of the egg. Furthermore, the choice of VD is shown to have a great effect, as can be seen in 4-24.



(a) (b) (c)

Figure 4-24: Schistosoma urine sample: The pixels found to be most similar to the endmembers found by the G4PN (a), G5PN (b) and G6PN (c) method as determined by the SAM metric

The same was found to be the case for the denoising algorithm and the choice of dimensionallity reduction, the ICA transform causing the method to find more pixel on the white blood cell to be endmembers. The PCA transform causes more pixels on the egg to be selected as endmembers. This can be seen in 4-25.



(a) (b) (c)

(d) (e) (f)

Figure 4-25: The locations of the pure pixels for the N5PN (a), G5PN (b), S5PN (c), N5IN (d), G5IN (e) and S5IN (f) methods on the Schistosoma sample highlighted by the yellow circles

In case of the NFINDR method, the spectral signatures are much more distinct from each other compared to the PPI method. They often contain one endmember that has high light transmittance in all wavelengths, thus likely corresponding to pixels showing the background. Furthermore, as can be seen in 4-26 the choice of denoising algorithm has little effect on the results,

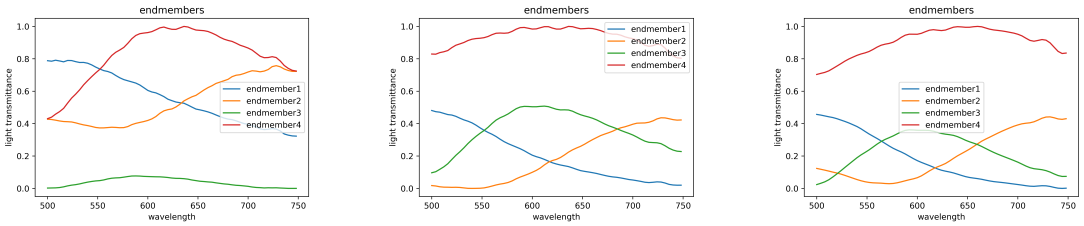


Figure 4-26: The spectral signatures of the endmembers found by the for the N6PN (a), G6PN (b), S6PN (c) methods on the Schistosoma sample

nor has the choice of dimensionallity reduction given a VD of 4 or 6 4-27.

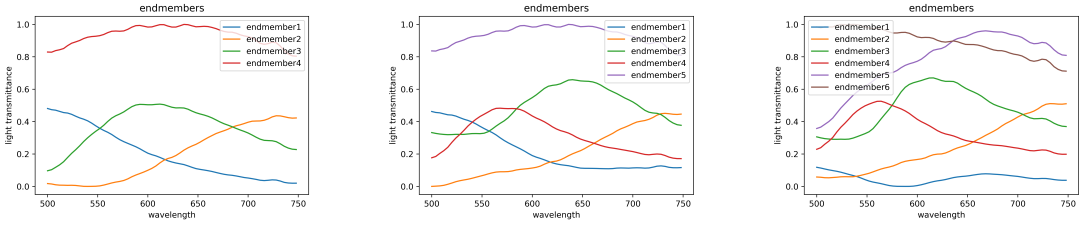


Figure 4-27: The spectral signatures of the endmembers found by the for the G4PN (a), G5PN (b), G6PN (c) methods on the Schistosoma sample

A VD of 6 was chosen as this resulted in the clearest abundance maps, as can be seen in 4-28. Combined with the PCA transform it is able to separate the egg from the background reasonably well, no matter the denoising algorithm.

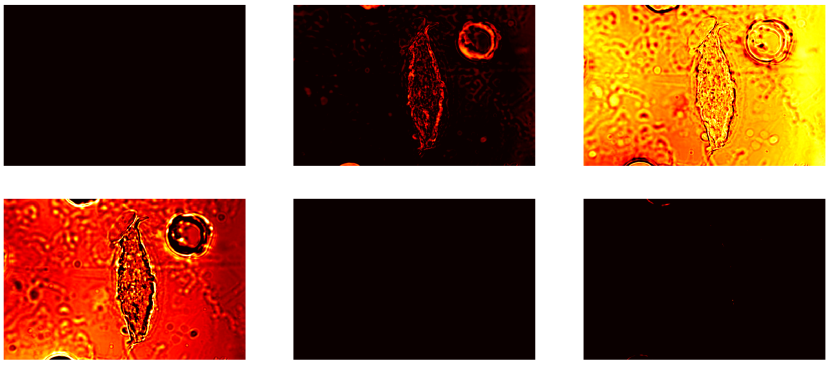


Figure 4-28: The endmembers derived in the S6PN method with negative light transmittance values

Statistics Based

The locations of the pixels most similar to the endmembers of the statistics based method differ greatly across all combinations of VD, dimensionallity reduction and denoising, some being shown in 4-29. The denoising, the 3d gaussian filtering cause the pixels corresponding to the endmembers to be mostly located on the white blood cell, while no or savitsky golay results in only one or two pixels on the white blood cell. The results across different virtual dimensionallities are much more consistent when the PCA transform is used compared to the ICA transform, where any change in VD causes a large change in the locations of the pixels corresponding to the endmembers. What is furthermore notable is that using the ICA transform, there often are no two endmembers with the same pixel as the most similar, neither are there neighbouring pixels most similar. This is different from when the method is applied to the stained Malaria sample.

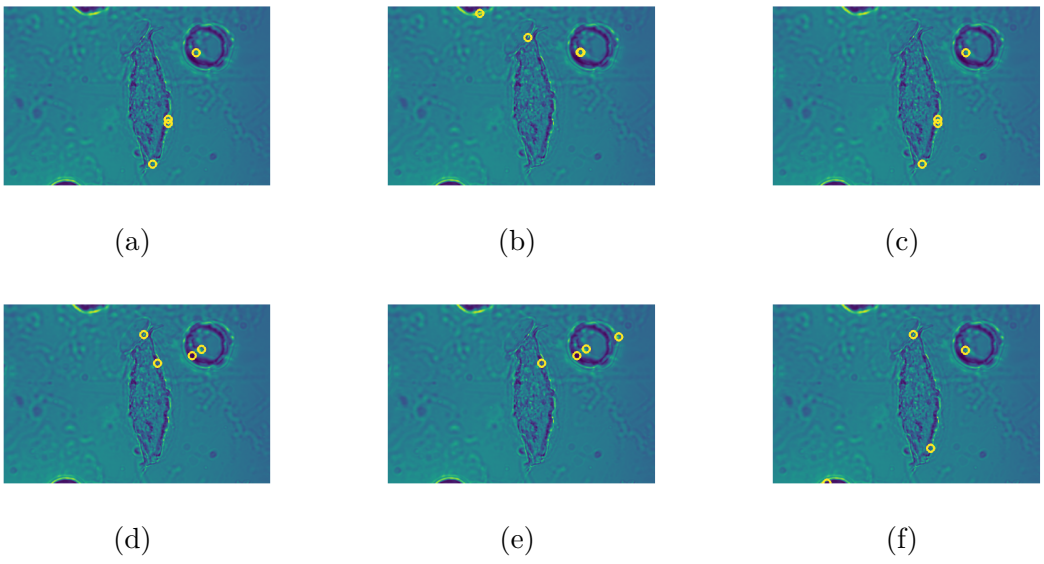


Figure 4-29: The locations of the pure pixels for the N4PB (a), G4PB (b), S4PB (c), N4IB (d), G4IB (e) and S4IB (f) methods on the Schistosoma sample highlighted by the yellow circles

The statistics based method shows similar results to the NFINDR method in that it shows spectral signatures being distinct from each other. Figure 4-31 shows that when coupled with the PCA transform the method is even able to find an endmember that could correspond to the background. This is not the case for the ICA or HySime transform, in which case an endmember with low light transmittance across all wavelengths is extracted, the exact opposite of the expected background spectrum.

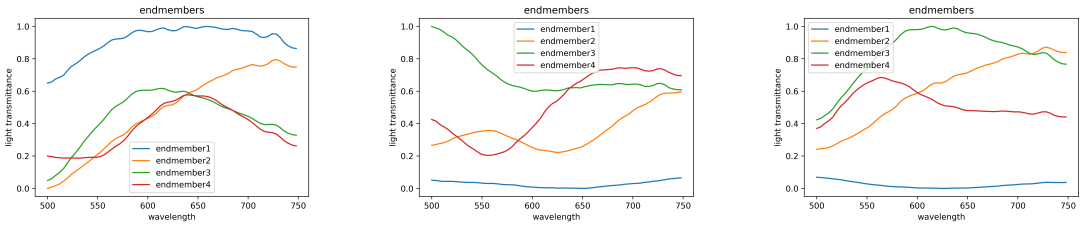


Figure 4-30: The spectral signatures of the endmembers found by the for the N4PB (a), N4IB (b), N4HB (c) methods on the Schistosoma sample

However, the best abundance maps seem to be derived when coupled with the HySime dimensionallity reduction, in which case one clearly shows the Schistosoma egg separated from the background quite well independent of the given VD, though again not from the white blood cell, as can be seen in 4-31.

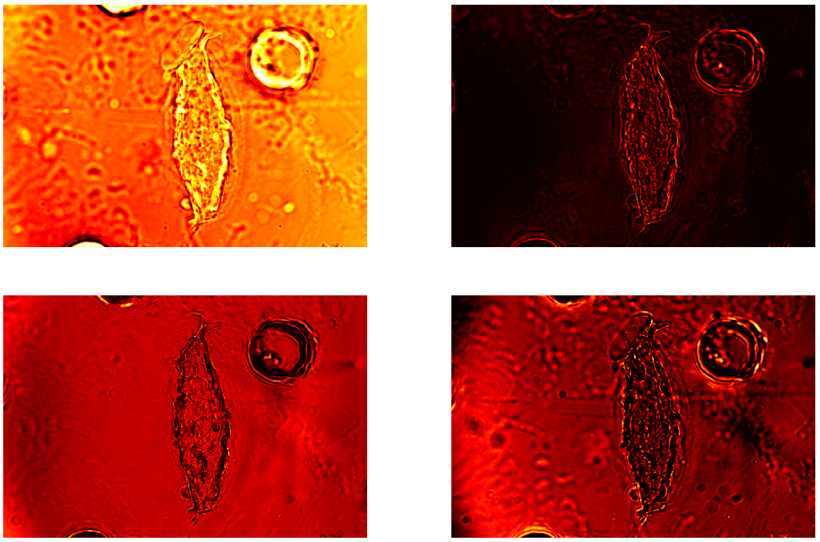
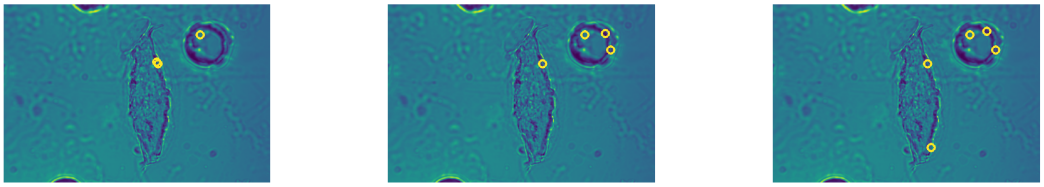


Figure 4-31: The endmembers derived in the S4HB method with negative light transmittance values

Sisal

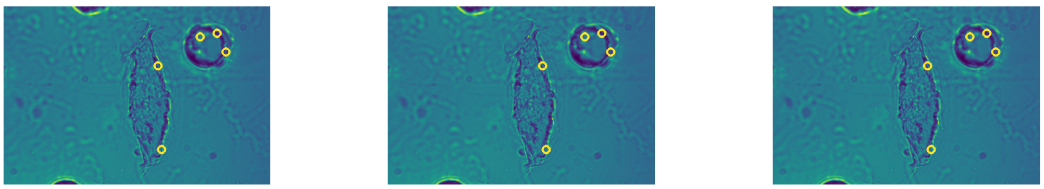
Finally, the locations of the most similar pixels to the endmembers given by the Sisal endmember extraction were found to be quite similar across different virtual dimensionallities, except for when a VD of 3 was given, as can be seen in 4-32 a to c,



(a) (b) (c)

Figure 4-32: Schistosoma urine sample: The pixels found to be most similar to the endmembers found by the G4S (a), G5S (b) and G6S (c) method as determined by the SAM metric

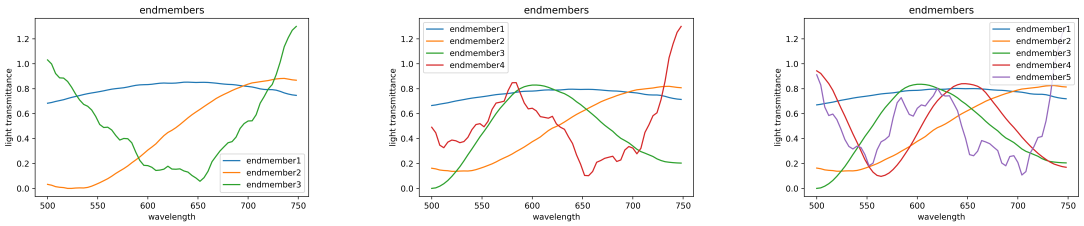
as well as denoising algorithms, as can be seen in 4-33 d to f. Notable is that most pixels most similar to the endmembers are located on the white blood cell and none directly on the background.



(d) (e) (f)

Figure 4-33: Schistosoma urine sample: The pixels found to be most similar to the endmembers found by the N5S (a), S5S (b) and G5S (c) method as determined by the SAM metric

The corresponding spectral signatures show each endmember being distinct from each other. It furthermore shows that each set contains an endmember with high values for all wavelengths, likely corresponding to the background. What is interesting is that given different virtual dimensionallities, endmember 1 and 2 remain similar, but the others change drastically, as can be seen in 4-34. The choice of denoising method has little to no effect as was the case with the Malaria sample.



(a) (b) (c)

Figure 4-34: The spectral signatures of the endmembers found by the for the G3S (a), G4S (b), G5S (c) methods on the Schistosoma sample highlighted by the yellow circles

However, none of the corresponding abundance maps showed the egg separate from the background. The closest it achieved was using no denoising and a VD of 6, as can be seen in 4-35, but it still falls short compared to the other methods in this regard.

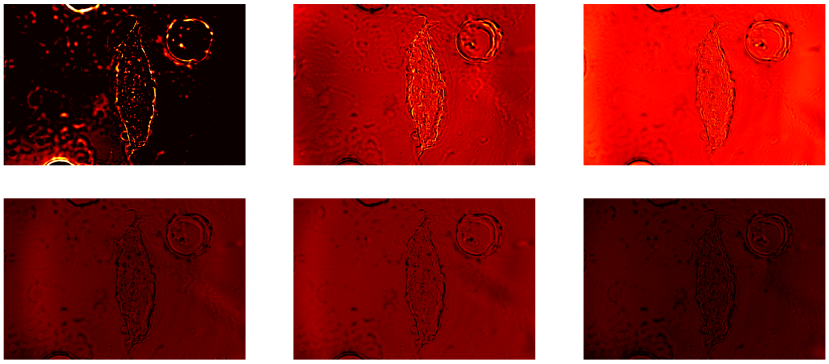


Figure 4-35: The endmembers derived in the N6S method with negative light transmittance values

4-3-4 Unstained Malaria Infected Blood Samples

The methods are similarly tested on the images of an unstained Malaria sample. By imaging the exact same location on the sample in stained and unstained condition it is possible to validate the results by referencing the images of the exact same location on the stained sample. The location of the parasites are thus known and can be similarly compared with the locations found by the method. The setup is the same as in the stained case with a 20x apochromatic magnification objective lens and a small cutout of the image containing no large distortions. The same range of wavelengths of 500 to 748 was used. The assumption was made that the cells and parasites would remain in place as the giemsa stain was applied. However, though they remain mostly in the same position, it is not exact as can be seen in 4-36. Compared to 4-37 all the cells moved to the left slightly (corrected for in the images), changed form slightly and some cells that were touching are no longer doing so. Since the two maps can thus not be aligned the semi-supervised endmember derivation discussed in 3-7 can not be applied. However, they have more or less retained their relative locations and it should still be possible to use the abundance map of the sample in stained condition to see which cells contains a parasite.

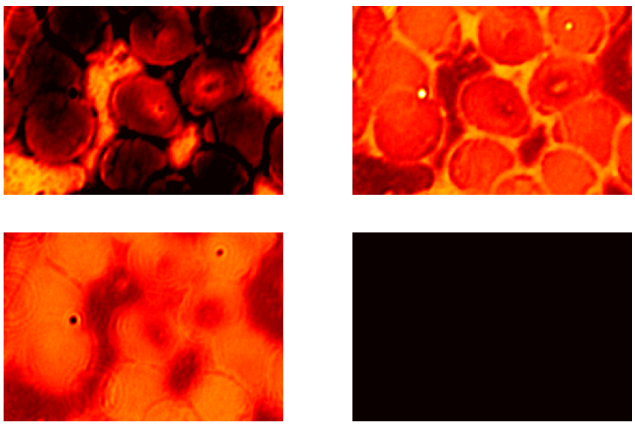


Figure 4-36: The abundance maps derived using the G4S method on the stained Malaria sample

These abundance maps were acquired using the the Sisal endmember extraction using 3d gaussian filtering and given a VD of 4. Though the abundance map is of lower quality than on the other sample, likely caused by the lower abundance of parasites, it still clearly shows a parasite in the upper middle of the image. Applying the endmember extraction directly on the data of the unstained parasite sample, no abundance map had higher intensity spots on these locations, as can be seen in below.

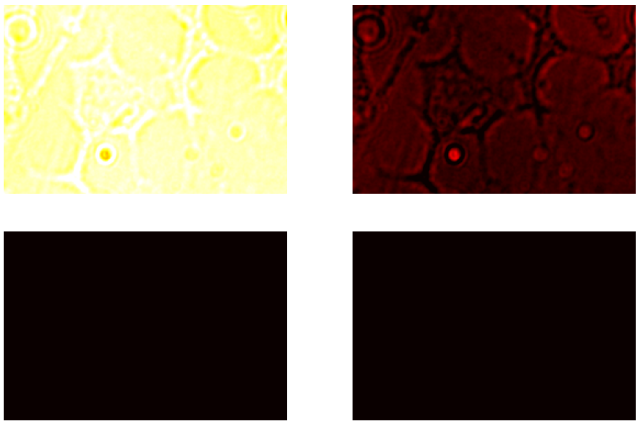


Figure 4-37: The abundance maps derived using the G4S method on the unstained Malaria sample

Due to the much lower contrast of the images of the unstained sample the endmember extraction methods mostly picks up on a small contamination on the microscope, the circular shape

to the left of the middle, and is not able to extract the spectral signature of the parasite. None of the other methods showed a brighter spot in the infected cell either.

4-4 Reference Spectrum Assisted Classification

Though the endmember extraction was unsuccessful for the unstained sample, in case of the stained Malaria parasite some of the endmembers were found to be reliable estimations of the spectral signature. Up until now these have been validated using the endmembers locations, spectra and abundance maps which can show clear errors, but give little to compare the seemingly well working methods by. Therefore, these spectral signatures are compared to each other in a similar way to how the red-green and principal component images were compared, namely classification. Furthermore, they are also compared to the classifeir using the first principal component of the hyperspectral and the simulated red-green (same as before) data. In order to derive a classification performance using the spectral signatures they are first used to create a detection map. From these detection maps the features will be derived using the OSP detection method, which uses all the determined endmembers, or the CEM detection method, which only requires the spectral signature of the parasite. The features derived from these detection maps are used to train a classifier which is to determine whether a cell is infected or not. The more accurate the spectral signatures are, the more the parasites will light up in the detection map, the better the classifier works. In this test the same set of features per image is used as before , namely mean, variance, dissimilarity and contrast. Furthermore, two types of classifier are used. The support vector machine with a regularization parameter of 1000 and a balanced class weight as well as the random forest classifier with 100 individual trees and a balanced class weight as well. The balanced class weight was chosen as in the sample set the ration of uninfected cells to infected cells is 5 to 1. The sensitivity, specificity and success rate of the methods that did well (ranked by their sensitivity using the SVM classifier) using OSP detection can be seen in 4-2.

	OSP-SVM	sens	spec	s	OSP-Forest	sens	spec	s
G4S		0.913	0.923	0.922		0.826	0.986	0.964
G3HB		0.870	0.910	0.903		0.826	0.986	0.958
S3HP		0.826	0.888	0.880		0.783	0.986	0.946
G5PN		0.739	0.832	0.819		0.739	0.965	0.892

Table 4-2: The sensitivity, specificity and success rate of the SVM and random forest classifier using the endmembers derived in the various endmember extraction methods in the OSP detection method.

Using the OSP detection, which uses all of the available endmembers, the endmembers determined in the Sisal method result in the highest sensitivity, specificity and success rate. The random forest classifier has higher success rate the SVM, but does so by sacrificing sensitivity, which leads to higher performance due to the relatively larger amount of negative samples. The results of the same test but using CEM detection are given in 4-3.

	CEM-SVM	sens	spec	p	CEM-Forest	sens	spec	p
G4S		0.913	0.881	0.885		0.652	0.990	0.940
G3HB		0.870	0.916	0.892		0.700	0.990	0.946
G5HN		0.870	0.909	0.887		0.610	0.990	0.933
N4HP		0.870	0.888	0.880		0.652	0.993	0.946

Table 4-3: The sensitivity, specificity and success rate of the SVM and random forest classifier using the endmembers derived in the various endmember extraction methods in the CEM detection method.

Using the CEM detection method, The highest success rate is actually produced using the endmember from statistics based. However, the highest sensitivity is still produced using the endmember corresponding to the Sisal method. Again, the random forest classifier has a higher succes rate, but at the cost of sensitivity. Finally, the results of these endmember assisted classification methods can be compared to the same algorithms trained on the first principal component, which requires no a priori knowledge, given in 4-4.

	sens	spec	p
PCA-SVM	0.740	0.867	0.850
PCA-Forest	0.522	0.993	0.930

Table 4-4: The sensitivity, specificity and success rate of the SVM and random forest classifier using the first principal component of the hyperspectral data

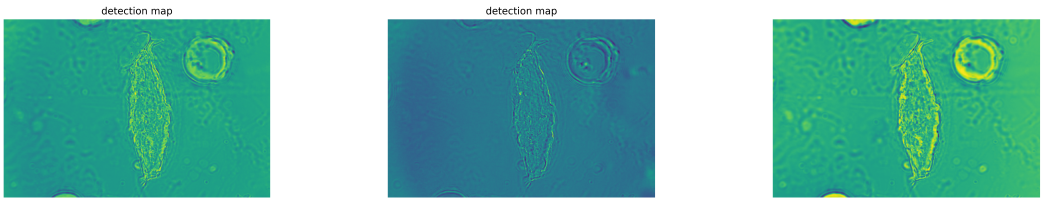
Which show us that a priori knowledge of the parasites spectral signature leads to significantly better classification performance which is missed when simply using PCA for dimensionallity reduction. Interestingly, apart from Sisal where the dimensionallity reduction is part of the method, the methods which seemingly give the best spectral signatures in terms of detection are those resulting from methods using the HySime dimensionallity reduction. The corresponding abundance maps, displayed in 4-38, generally show the background and foreground separated really well, but the RBC's and parasites not.

(a) (b)

Figure 4-38: The abundance maps resulting from the S3HP (a) and G3HB (b) algorithm

4-5 Reference Spectrum Assisted Detection

In case of the Schistosoma eggs it is a bit more difficult. Current methods of detection generally involve some kind of convolutional neural network, which are very heavy to train, requiring many training samples. At this time training a similar network is not an option due to the limited samplesize of hyperspectral images. Therefore, the spectral signatures are once again used to create detection maps using OSP and CEM detection. Out of these maps the ones that show the egg the most clearly are compared to the first principal component which requires no a priori knowledge. These images are displayed in 4-39.



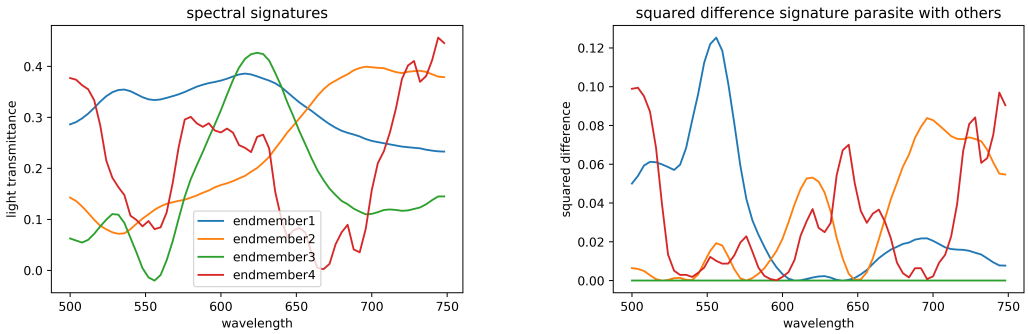
(a) (b) (c)

Figure 4-39: The resulting detection maps using OSP detection with the endmembers of the N4IP algorithm (a), CEM detection with the N6HB algorithm (b) and the first principal component (c).

Here, the first principal component, which does not require the spectral signature of the egg to be known, displays the it similarly if not more clearly than the others. Unlike the abundance maps, the detection methods do not provide an image which make the eggs more easily detectable.

4-6 Comparing Multispectral and bright field RGB Classification

Up until now some spectral signatures have been found to be a reliable estimation of the true spectrum of the giemsa stained Malaria parasite. It has also been shown that when using these spectral signatures in the feature derivation hyperspectral imaging can be used to achieve superior classification performance. However, It is not realistic to expect labs in sub-saharan Africa to be in the possession of a hyperspectral microscope in order to diagnose Malaria. Therefore, the next step is to determine whether the derived spectral signatures can be used in the development of methods using more cost-effective equipment. The method that is explored is that of a multispectral microscopy. Using the newfound spectral signatures the most discriminative wavelengths can be determined, the wavelengths where the signature of the parasite is has the largest squared difference in light transmittance to the other endmembers, as displayed in 4-40.



(a) (b)

Figure 4-40: The resulting spectral signatures from the endmember extraction (uncorrected for illumination) (a) and root mean square difference between the spectral signature corresponding to the parasite and the others (b)

Imaging at just the two wavelengths where the differences are the largest, in this case 552nm and 692nm, should give an advantage over using the red-green images. To test this two sets of two images were derived from the same data. The images in the first set are made by averaging the 5 images at the discriminative wavelength and nearest neighbouring wavelengths. The second set again aims to represent the red and green wavebands of an RGB image, but this time only 5 images equidistant in the relevant spectral domain are used. All the images used for classification are derived from 5 input images should thus all have a comparable signal-to-noise ratio while keeping the resolution at these specific wavelength high enough to take advantage of the increased discriminative power. The results of the SVM algorithm can be seen in 4-5

	sens	spec	s
DW-SVM	0.826	0.930	0.915
RG-SVM	0.739	0.867	0.849

Table 4-5: The sensitivity, specificity and success rate of the SVM classifier using the discriminative wavelength images versus the red-green images

and the random forest classifier in 4-6.

	sens	spec	s
DW-Forest	0.739	0.979	0.945
RG-Forest	0.565	0.979	0.921

Table 4-6: The sensitivity, specificity and success rate of the random forest classifier using the discriminative wavelength images versus the red-green images

The results clearly show the classifiers trained on the discriminative wavelength images outperforming those trained on the red-green images both in the SVM and the random forest classifier. This implies that such a multispectral setup could indeed give an advantage compared to RGB imaging, provided this result holds when the blue waveband is included. Using

more wavebands could improve the performance further, but in the field this would translate to a slower data acquisition process.

Chapter 5

Discussion and Conclusion

The results shown in the previous chapter have shown some of the strengths and weaknesses of hyperspectral imaging in the application of Malaria detection in stained and unstained thin smear blood samples and Schistosoma egg detection in urine samples. This chapter will discuss each of the tested methods in relation to the research questions. It will then give a conclusion on these questions and the potential of hyper- and multispectral imaging for this application as a whole as well as give a recommendation for possible future directions of related research.

5-1 Discussion

A great many combinations of virtual dimensionallity, denoising method, dimensionallity reduction and endmember extraction have been applied to the problem of Malaria and Schistosoma detection, with various success. The effectiveness of these methods will be discussed per type of sample.

5-1-1 The Stained Malaria Sample

Arguably the easiest of the tasks is the spectral signature estimation of the stained Malaria infected thin smear blood sample. It has been possible to detect Malaria prasites in these kinds of samples for some time. Traditionally, light microscopy with an RGB sensor combined with some RBC segmentation methods and various classifiers have been applied to this problem, often successfully. The application of hyperspectral microscopy on the stained Malaria sample is therefore mostly focused on the estimation of the spectral signature of both the stained parasite and the other substances in the sample. This has been done with the intention of determining the wavelengths in which this spectral signature has the largest difference to the other endmembers in terms of light transmittance and has been shown to be possible to various degrees.

Endmember Extraction

The PPI method is shown to be inconsistent, having different results each run, making it an unreliable tool for spectral signature estimation. The NFINDR method seems to perform well looking at the spectra, endmember location and abundance maps, but has a low corresponding classification performance. Statistics based is unable to separate the RBC's and background from each other. The Sisal method performs the best of the tested methods. Both the locations of the most similar pixel as well as the shape of the endmembers in the spectral domain are found to be in line with expectation. The abundance maps show the RBC's, background and parasites separated quite well and the endmembers resulting from the Sisal method are in turn found to correspond to the best classification performance. When OSP is used to create the detection map the classifier using the corresponding endmembers has higher sensitivity and specificity than the classifiers trained on other sets of endmembers. It also outperforms the classifiers trained on the first principal component of the data or the red-green images. When CEM is used to create the detection maps from which the features are derived the performances are more similar, but the highest sensitivity is still achieved using the endmembers derived using Sisal. An interesting finding in the classification tests is that when using OSP detection the other endmember extraction methods methods are often best coupled with the HySime dimensionallity reduction. These often correspond to abundance maps in which the back- and foreground are separated well, but do not necessarily separate the RBC's and the parasites. This could be implying that during the OSP detection valuable information is lost in the removal of the endmembers which correspond to the RBC's as compared to only removing the background. This would have to be researched further.

Multispectral Classification

The use of these estimated spectral signatures for the determination of the most discriminative wavelengths proved to be fruitful. The use of images at the most discriminative wavelengths, the wavelengths where the spectral signature of the parasite has the largest difference to the other endmembers in terms of light transmittance, leads to significantly higher sensitivity and specificity compared to using the simulated RGB data. This is shown to be the case the case using either the SVM classifier or the random forest classifier. Of course, the RGB data is simulated from the hyperspectral data and made to have a similar signal-to-noise ratio as the specific wavelength images, which is not representative to the real world where the RGB images have a much stronger lightsource. However, the results are still promising and further research should be done to determine whether this could have a real world use.

5-1-2 The Schistosoma Sample

The next task is the estimation of the spectral signature of the Schistosoma eggs. As the eggs are mostly transparent when using conventional bright field microscopy, it often requires powerful convolutional neural networks in order to be able to detect them. The application of hyperspectral microscopy is therefore focused on creating an image on which the eggs are more easily detected. The abundance maps which show the Schistosoma egg most clearly are in the occasional successful PPI endmember extraction run. Similarly to before, the method is not very reliable, resulting in different spectra in multiple runs of the same algorithm. NFINDR

and statistics based has been shown to be much more reliable in this regard, providing an abundance map where the Schistosoma egg is almost as visible as in case of the PPI algorithm, but doing so each time the algorithm is run. The Sisal method is not able to create an abundance map in which the egg is clearly visible. Finally, though the goal of more easily detectable eggs was sometimes achieved using FCLS regression, none of the corresponding detection maps, using the computationally lighter OSP or CEM detection, provided the same benefit when compared to the first principal component image.

5-1-3 The Unstained Malaria Sample

None of the endmember extraction algorithms is able to find the spectral signature of the unstained Malaria parasite. The endmember locations are not positioned on the parasites, nor does one of the abundance maps have the parasites specifically light up. An interesting finding is that the cells change shape and position slightly in the process of staining. This makes the use of spatial information from the sample in stained condition for the analysis of the sample in unstained condition more difficult than expected.

5-2 Conclusion

In search of the potential advantages that multi- and hyperspectral imaging can provide first the classification of the RBC's as being infected or not was tested using the hyperspectral data directly versus the RGB data simulated from the same dataset. Interestingly, the classifier using the full hyperspectral dataset is shown to be worse than using the RGB data in terms of sensitivity and specificity. However, as later test do show increased performance using a different approach, this is most likely due to the the small samplesize. It is not uncommon in machine learning that, given additional data, the classification performance remains similar or even decreases if not combined with sufficient training data. To prevent this, a similar test using a much larger sample size and additional features should be conducted. For this an altered setup in which the imaging is automated and/or the pushbroom scanning method is used would be required to handle this task as the process is too slow using the setup used for this study.

In case of the stained Malaria sample, of all the endmember extraction methods, with the various combinations of denoising, VD and dimensionallity reduction, The Sisal method using 3d gaussian filtering is found to result in the most plausible spectral signatures. The endmembers themselves, the locations and the abundance maps are all in line with expectations. Furthermore, the classification which used the signatures derived from the Sisal method in the feature derivation process have higher sensitivity and specificity compared not only to the other endmember extraction methods, but also to both the RGB simulated data and the first principal component derived from the full hyperspectral data. This shows that hyperspectral imaging can indeed provide an significant advantage over RGB imaging in this application. It furthermore implies that additional information has been acquired in the endmember extraction process and that the derived spectral signatures can be considered reasonably accurate.

As for the challenge of spectral signature estimation for the Schistosoma eggs, this proved to be more difficult. All endmember extraction methods but Sisal are able to separate the egg

from the background given a high enough VD, but not able to separate the egg and the white blood cell from each other. Using FCLS regression using the spectral signatures derived in the endmember extraction, it is often possible to create an abundance map in which the eggs are clearly separated from the background. This provides a significant advantage over the RGB representation and the first principle component, but the process is computationally heavy. Using the computationally lighter OSP or CEM target detection with the estimated signature corresponding to the egg does not provide the same advantage. It therefore remains unclear whether an accurate prediction of the spectral signature of the Schistosoma egg is made. A similar classification comparison as is done with the stained Malaria sample was not possible in this study. This is because Schistosoma egg detection is currently mainly done using comvolutional neural networks which require a lot of training data. Here the small samplesize proves to be the limiting factor. However, this should be considered as a direction of future research. As stated before, this would require the imaging process to be significantly quicker.

Finally, none of the tested methods are able to extract the spectral signature of the unstained Malaria parasite when applying the endmember extraction directly on the data. This likely means the spectral signature is simply not different enough from the RBC's at the tested wavelengths and the endmember extraction methods are not able to make an accurate estimation of the spectral signature of the parasite. Perhaps a stronger light source could be used, making a higher spectral resolution and lower signal-to-noise ration possible, but trying to use different part of the electromagnetic spectrum is likely to be more fruitful.

The proposed hypothetical multispectral microscope which images just at the most discriminative wavelengths is shown to increase performance compared to its RGB counterpart when both are derived from the same data. However, it has to be considered that this comparison uses images which are purposefully made to have a similar signal-to-noise ratio, while normally bright field microscopy would provide a much higher signal-to-noise ratio due to the stronger illumination. Further research is required to determine whether multispectral imaging could provide a real world advantage.

Bibliography

[1] C. van Engelenburg, G. Vdovin, and T. Agbana, "Potential enrichments in malaria diagnostics : hyperspectral imaging and group-equivariant neural networks (Thesis work)," *master thesis*, 2020.

[2] M. Poostchi, K. Silamut, R. J. Maude, S. Jaeger, and G. Thoma, "Image analysis and machine learning for detecting malaria," *Translational Research*, vol. 194, pp. 36–55, 2018.

[3] wikiwand, "Lens." https://www.wikiwand.com/en/Refraction. (accessed: 06.10.2021).

[4] wikipedia, "Lens." https://en.wikipedia.org/wiki/Lens. (accessed: 06.10.2021).

[5] S. Doyle, "Examination of blood for parasites." https://slideplayer.com/slide/5762715/. (accessed: 31.05.2021).

[6] scikits learn, "Histogram equalization." https://scikit-image.org/docs/0.5/auto$_e$*xamples/plot*$_e$*qualize.html*. (*accessed* : 31.05.2021).

[7] G. Moallem, H. Sari-Sarraf, M. Poostchi, R. J. Maude, K. Silamut, S. Antani, G. Thoma, S. Jaeger, and M. Amir Hossain, "Detecting and segmenting overlapping red blood cells in microscopic images of thin blood smears," no. March 2018, p. 50, 2018.

[8] J. S., "Malaria datasets." https://lhncbc.nlm.nih.gov/LHC-publications/pubs/MalariaDatasets.html. (accessed: 31.05.2021).

[9] M. A. Shahin and S. J. Symons, "Detection of hard vitreous and starchy kernels in amber durum wheat samples using hyperspectral imaging (GRL Number M306)," no. March, 2015.

[10] C. I. Chang, *Hyperspectral Data Processing: Algorithm Design and Analysis*. 2013.

[11] J. M. Bioucas-Dias, A. Plaza, N. Dobigeon, M. Parente, Q. Du, P. Gader, and J. Chanussot, "Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 5, no. 2, pp. 354–379, 2012.

[12] juliaimages, "grey level co-occurence matrix." https://juliaimages.org/ImageFeatures.jl/stable/tutorials/glcm/. (accessed: 06.01.2021).

[13] C. Cortes and V. Vapnik, "Support-Vector Networks," vol. 297, pp. 273–297, 1995.

[14] M. Cavaoini, "Machine learning: Decision tree classifier." https://medium.com/machine-learning-bites/machine-learning-decision-tree-classifier-9eb67cad263e. (accessed: 07.09.2021).

[15] Wikipedia, "Apochromat." https://en.wikipedia.org/wiki/Apochromat. (accessed: 13.10.2021).

[16] Global Malaria Programme: WHO Global, *World malaria report 2019*. No. December, 2019.

[17] B. Ngasala, M. Mubi, M. Warsame, M. G. Petzold, A. Y. Massele, L. L. Gustafsson, G. Tomson, Z. Premji, and A. Bjorkman, "Impact of training in clinical and microscopy diagnosis of childhood malaria on antimalarial drug prescription and health outcome at primary health care level in Tanzania: A randomized controlled trial," *Malaria Journal*, vol. 7, pp. 1–11, 2008.

[18] D. Payne, "Use and limitations of light microscopy for diagnosing malaria at the primary health care level," *Bulletin of the World Health Organization*, vol. 66, no. 5, pp. 621–626, 1988.

[19] C. K. Murray, R. A. Gasser, A. J. Magill, and R. S. Miller, "Update on rapid diagnostic testing for malaria," *Clinical Microbiology Reviews*, vol. 21, no. 1, pp. 97–110, 2008.

[20] A. Moody, "Rapid diagnostic tests for malaria parasites," *Clinical Microbiology Reviews*, vol. 15, no. 1, pp. 66–78, 2002.

[21] CDC, "Malaria - biology." https://www.cdc.gov/malaria/about/biology/index.html. (accessed: 21.12.2020).

[22] S. S. Ranhotra, "An Alternative Approach to Detect the Presence of Schistosoma Haematobium Infection in Affected Regions of Benue State-Nigeria," pp. 2113–2117, 2017.

[23] K. E. Pe and E. A. Villacorte, "Automated Detection of Helminth Eggs in Stool Samples Using Convolutional Neural Networks," pp. 750–755, 2020.

[24] D. C. Warhurst and J. E. Williams, "ACP Broadsheet no 148. July 1996. Laboratory diagnosis of malaria.," *Journal of Clinical Pathology*, vol. 49, no. 7, pp. 533–538, 1996.

[25] S. M. Parsel, S. A. Gustafson, E. Friedlander, A. A. Shnyra, A. J. Adegbulu, Y. Liu, N. M. Parrish, S. A. Jamal, E. Lofthus, L. Ayuk, C. Awasom, C. J. Henry, and C. P. McArthur, "Malaria over-diagnosis in Cameroon: Diagnostic accuracy of Fluorescence and Staining Technologies (FAST) Malaria Stain and LED microscopy versus Giemsa and bright field microscopy validated by polymerase chain reaction," *Infectious Diseases of Poverty*, vol. 6, no. 1, p. 1, 2017.

[26] S. Moon, S. Lee, H. Kim, L. H. Freitas-Junior, M. Kang, L. Ayong, and M. A. Hansen, "An Image Analysis Algorithm for Malaria Parasite Stage Classification and Viability Quantification," *PLoS ONE*, vol. 8, no. 4, 2013.

[27] Nobuyuki Otsu, "A Threshold Selection Method from Gray-Level Histograms," *IEEE Trans. Syst. Man Cybern*, vol. 9, no. 1, pp. 62–66, 1979.

[28] S. W. Sio, W. Sun, S. Kumar, W. Z. Bin, S. S. Tan, S. H. Ong, H. Kikuchi, Y. Oshima, and K. S. Tan, "MalariaCount: An image analysis-based program for the accurate determination of parasitemia," *Journal of Microbiological Methods*, vol. 68, no. 1, pp. 11–18, 2007.

[29] Bellman, *Dynamic programming*. Dover Publications, 2003.

[30] R. M. Haralick, I. Dinstein, and K. Shanmugam, "Textural Features for Image Classification," *IEEE Transactions on Systems, Man and Cybernetics*, vol. SMC-3, no. 6, pp. 610–621, 1973.

[31] Z. Liang, A. Powell, I. Ersoy, M. Poostchi, K. Silamut, K. Palaniappan, P. Guo, M. A. Hossain, A. Sameer, R. J. Maude, J. X. Huang, S. Jaeger, and G. Thoma, "CNN-based image analysis for malaria diagnosis," *Proceedings - 2016 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2016*, pp. 493–496, 2017.

[32] M. J. E. Savitzky, A.; Golay, "Smoothing and Differentiation," *Anal. Chem*, vol. 36, no. 8, pp. 1627–1639, 1964.

[33] J. M. Bioucas-Dias and J. M. Nascimento, "Hyperspectral subspace identification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 46, no. 8, pp. 2435–2445, 2008.

[34] J. Wang and C. I. Chang, "Independent component analysis-based dimensionality reduction with applications in hyperspectral image analysis," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 44, no. 6, pp. 1586–1600, 2006.

[35] C. Jutten and J. Herault, "Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture," *Signal Processing*, vol. 24, no. 1, pp. 1–10, 1991.

[36] A. Hyvärinen and E. Oja, "A Fast Fixed-Point Algorithm for Independent Component Analysis," *Neural Computation*, vol. 9, no. 7, pp. 1483–1492, 1997.

[37] Winter, "N-FINDR: an algorithm for fast autonomous spectral end-member determination in hyperspectral data," *Proceedings of SPIE - The International Society for Optical Engineering*, 1999.

[38] C.-i. Chang, S. Member, C.-c. Wu, and S. Member, "A New Growing Method for Simplex-Based Endmember Extraction Algorithm," no. November, 2016.

[39] M. D. Craig, "Minimum-Volume Transforms for Remotely Sensed Data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 32, no. 3, pp. 542–552, 1994.

[40] J. M. Bioucas-Dias, "A variable splitting augmented Lagrangian approach to linear spectral unmixing," *WHISPERS '09 - 1st Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing*, pp. 2–5, 2009.

[41] D. C. Heinz, S. Member, C. Chang, and S. Member, "Mixture Analysis Method for Material Quantification in Hyperspectral Imagery," *Analysis*, vol. 39, no. 3, pp. 529–545, 2001.

[42] C. M. Wang, C. C. C. Chen, Y. N. Chung, S. C. Yang, P. C. Chung, C. W. Yang, and C. I. Chang, "Detection of spectral signatures in multispectral MR images for classification," *IEEE Transactions on Medical Imaging*, vol. 22, no. 1, pp. 50–61, 2003.

[43] J. C. Harsanyi and C. I. Chang, "Hyperspectral Image Classification and Dimensionality Reduction: An Orthogonal Subspace Projection Approach," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 32, no. 4, pp. 779–785, 1994.

[44] L. E. O. Breiman, "Random Forests," pp. 5–32, 2001.