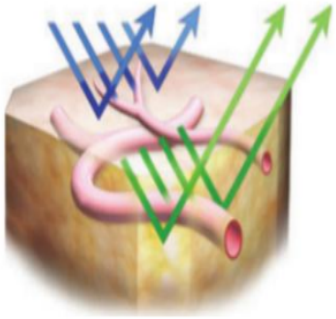


Narrow-Band Image Processing for Gastroenterological Examinations

Bruno Miguel Ferreira Mendes
Mestrado em Física Médica
Departamento de Física

Orientador
Ricardo Sousa, Doutor, Instituto de Telecomunicações - UP

Orientador
Carla Rosa, Professora Doutora, INESC TEC - UOSE, FCUP

Todas as correções determinadas pelo júri, e só essas, foram efetuadas.

O Presidente do Júri,

Porto, _____/_____/_____

To my parents...

"The Devil is in the Details"

Abstract

Narrow-band Imaging (NBI) is a recent and promising technique which is being applied to modern endoscopes. It allows to enhance the contrast between superficial and deeper vessels by illuminating the tissue with white light and having two filters in the Charged-Coupled Device (CCD) sensor with different wavelengths. These wavelengths match the absorption peaks of haemoglobin. Based on the different penetration depths of light (longer wavelengths penetrate deeper) the superficial vessels will be enhanced by blue and the deeper vessels will be enhanced by green. This increase in contrast allows a better identification of vascular alterations indicative of a pathology. It also brings new patterns that need to be interpreted in order to perform a correct and precise classification of these new images that contain information that is of difficult perception when using conventional white light. The special conditions that these new endoscopic images are acquired allows us to modulate these images with a physical model that describes the distribution of light in the tissue. With this in mind we can rebuild the information from the RGB channels and extract features that exhibit special photometric invariances.

In this thesis we built a Computer Aided Diagnosis (CAD) support system specialized to learn these new patterns in order to perform a correct and precise classification. For our system we developed a framework encompassing three standard steps: feature extraction and description, and pathology learning. A physical model for feature description of coloured gastroenterology images was assessed. To test the developed framework we used an image dataset with 250 endoscopic images from the oesophagus were 61 are normal and 189 present pre-cancer lesions. Converting the images to gray we obtained a performance of 79% and adding colour information we obtained a performance of 84% using the opponent colours.

Resumo

Narrow-band Imaging (NBI) é uma técnica recente e promissora, que tem sido aplicada a endoscópios modernos. Permite aumentar o contraste entre os vasos superficiais e mais profundos, iluminando o tecido com luz branca e colocando dois filtros no sensor Charged-Coupled Device (CCD) com diferentes comprimentos de onda. Estes comprimentos de onda correspondem aos picos de absorção da hemoglobina. Com base nas diferentes profundidades de penetração da luz (comprimentos de onda mais longos penetram mais profundamente), os vasos superficiais serão enaltecidos pelo azul e os vasos mais profundos pelo verde. Este aumento no contraste permite uma melhor identificação de alterações vasculares indicativas de patologia. Mas também mostram novos padrões que precisam de ser interpretados de forma a realizar uma classificação correta e precisa destas novas imagens que contêm informação que é de difícil percepção pelo uso de luz branca convencional. As condições especiais em que estas novas imagens endoscópicas são adquiridos permite-nos modulá-las com um modelo físico que descreve a distribuição da luz no tecido. Podemos então reconstruir as informações dos canais RGB e extrair características que apresentam invariâncias fotométricas especiais.

Nesta tese, construímos um Sistema Computorizado de Auxílio ao Diagnóstico (CAD) especializado para estes novos padrões por forma a realizar uma classificação correta e precisa. Para o nosso sistema, desenvolvemos uma abordagem que engloba três etapas: extração de características e descrição e aprendizagem da patologia. Um modelo físico para a descrição de características de imagens coloridas em gastroenterologia foi avaliado. Para testar o modelo desenvolvido foi utilizado um conjunto com 250 imagens endoscópicas do esófago em que 61 são normais e 189 apresentam lesões pré-cancerosas. Convertendo as imagens para cinza obtivemos um desempenho de 79 % e adicionando informação da cor obtivemos um desempenho de 84 %, usando as cores oponentes.

Acknowledgments

This thesis was a very challenging task. To enter in the world of Pattern Recognition and Machine Learning was not easy. I must confess that sometimes I felt a little lost trying to understand some of the terminologies and methods used in this field of Image Processing. I was able to overcome these difficulties thanks to my supervisor Doctor Ricardo Sousa. Thanks for being always available to answer my questions, for the patience in those moments I was struggling and making mistakes and for the great support given in this journey.

I also would like to thank to my supervisor Professor Doctor Carla Rosa for the suggestions and support given in the Physics formulation of this thesis and for providing a working space at INESC-Porto-UOSE.

A big thanks also to Universidade do Porto (Faculdade de Ciências) and to INESC-Porto.

To my parents, for the understanding and support despite of the difficulties that life brings. To my brother, that was my travelling comrade and had to put up with me every mornings, to all my friends for being there when I needed and last but not least, to my girlfriend, for the patience, love and support in this tough journey of my life.

The work developed in this thesis was financially supported by FCT (Portuguese Science Foundation) grant PTDC/EIA-CCO/109982/2009 (CAGE).

Contents

List of Tables

List of Figures

List of Algorithms

Chapter 1

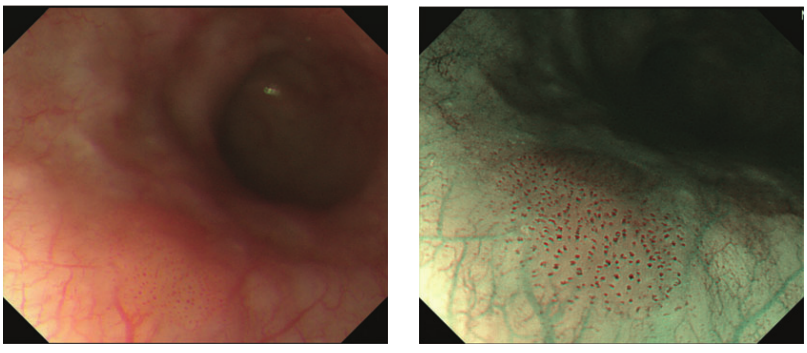
Introduction

"A picture is worth more than a thousand words". This old adage referring to the amount of information an image can convey has been around us for ages. Of course its origins had nothing to do with medical images, but its essence is totally applied. With the advances in science and technology, we are nowadays capable to obtain better quality images of the human body. This has brought major advantages in diagnosing, monitoring or treating a disease. In Medical Imaging it is common to use either invisible light, as in an X-Ray exam or visible light like in an endoscopic procedure. In both the interpretation of these images is drawn by the physicians experience and expertise adding always a subjective element to the analysis. With the introduction of computers in Medical Imaging the subjectiveness in the analysis has diminished. It was in this context that the Computer Aided Diagnosis (CAD) support systems appeared, to help in the decision and classification process. The advantage of having a system insensitive to fatigue or distraction is a major plus because we are also withdrawing the ambiguity and subjectiveness of human analysis. The first CAD systems appeared initially connected to endoscopic images (Liedlgruber & Uhl, 2011).

Endoscopy is a technique widely used in modern medicine to observe the inner cavities of the human body. It makes use of an endoscope which is basically a flexible tube that consists of a bundle of optical fibres. The endoscope has evolved a lot since the first rigid endoscope introduced in a demonstration of gastroscopy by Adolph Kussmaul in 1868. The modern ones allow to view a real time image on a monitor. They are called video endoscopes

and use a Charged-Coupled Device (CCD) for image generation. A CCD chip is an array of individual photo cells (or pixels) that receive photons reflected from a surface and produce electrons proportionally to the amount of light received. This information is then stored in memory chips and processed in a monitor (Muto et al., 2009).

A recent technique in endoscopy consists in using only certain wavelengths of visible light by placing a filter in front of the light source therefore narrowing the bandwidth of the light output. This technique is called Narrow-band Imaging (NBI) and is a very promising tool in the diagnosis of gastrointestinal diseases. The NBI system uses two specific wavelengths, 415 nm and 540 nm that match the absorption peaks of haemoglobin. NBI can be used in a RGB Sequential System which consists of inserting a RGB rotary filter in front of the light source but only the green and blue filter are activated. Another approach is to place colour filters in each pixel of the CCD chip. In both systems a Xenon lamp is used as a light source (Muto et al., 2009). In the illumination process of the tissue using NBI, blue is mainly absorbed by superficial vessels while green continues to penetrate the tissue and is absorbed by deeper vessels (see the absorption spectra of haemoglobin in Figure 2.3). The capillaries in the superficial mucosal layer are then emphasized by the blue light and the deeper mucosal and submucosal vessels are made visible by the green light. However, to reproduce a colour image in the monitor we need three images in the R, B and G channels. The R channel records the signal derived from green illumination so the vessels in the deeper layer will have a cyan colour. The B and G channels record the signal derived from blue illumination and the superficial vessels will appear brownish (Muto et al., 2009).



(a) White Light. (b) NBI filter.

Figure 1.1: Effect of NBI (Figure from Muto et al. (2009))[1].

[1]Better viewed in colour.

Figure 1.1 shows how a Human sees a tissue without (left figure) and with NBI (right figure). In the normal image vascular patterns are difficult to visualize. The surface of the oesophagus appears smooth. Depending of the NBI technology employed, we are able to identify some polyps and a vascular pattern at the surface with a brownish colour. Deeper vessels are also emphasized appearing with a slight cyan colour. Despite of the technology, NBI is a very promising tool in the early diagnosis of gastroenterological pathologies decreasing the examination time, reducing unnecessary biopsies and increasing the accuracy of such examinations (Muto et al., 2009).

1.1 Motivation

In the twentieth century the average life expectancy from birth in Portugal has increased and in 2011 was of about 80 years. The main causes of death are cardiovascular diseases and cancer[2]. Although the efforts for prevention and early detection have been made, cancer is still an issue in public health. According to *Registo Oncológico Regional do Norte* (RORENO) the total number of new patients observed at *Instituto Português de Oncologia do Porto Francisco Gentil* (IPO-Porto) in 2010 was of 10241[3]. From those 7050 were malignant. The most frequent oncological pathology observed was the gastrointestinal tract, 23,2% of the cases, followed by the genitourinary organs, 22,2%, and breast with 19,9%[3]. From the diagnosed cancers within the gastrointestinal tract one of the deadliest is the oesophageal cancer together with the liver and pancreas.

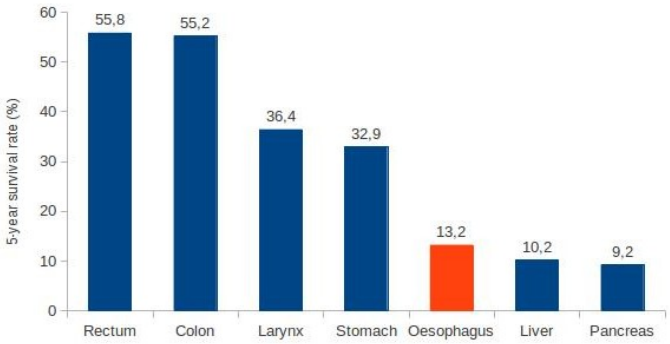


Figure 1.2: 5-year survival rate for some parts of the gastrointestinal tract[3].

[2]Data available at http://www.pordata.pt/Portugal
[3]Data available at http://www.roreno.com.pt/

Oesophageal cancer is classified mainly in two groups: squamous cell carcinoma and adenocarcinoma. According to RORENO the squamous cell carcinoma represents 101 of the 121 diagnosed cases and adenocarcinoma represents 14 cases and other tumours with 6 cases[3].

The disease can also be classified in terms of its extension in five groups[4].

In Situ	malignant tumour that has not penetrated the basement membrane not extended beyond the epithelial tissue
Localized	invasive malignant tumour confined to the organ of origin
Regional	malignant tumour that has extended beyond the limits of the organ of origin directly into surrounding organs and tissues or evolved through the lymphatic system or both
Distant	malignant tumour that has spread to parts of the body remote from the primary tumour either by direct extension or by metastasis
Not Recorded	insufficient information to assign a stage

Table 1.1: Classification of Oesophageal cancer according to its extension.

Most of oesophageal cancer is regional or distant representing 51 and 35 respectively of the 121 diagnosed cases.

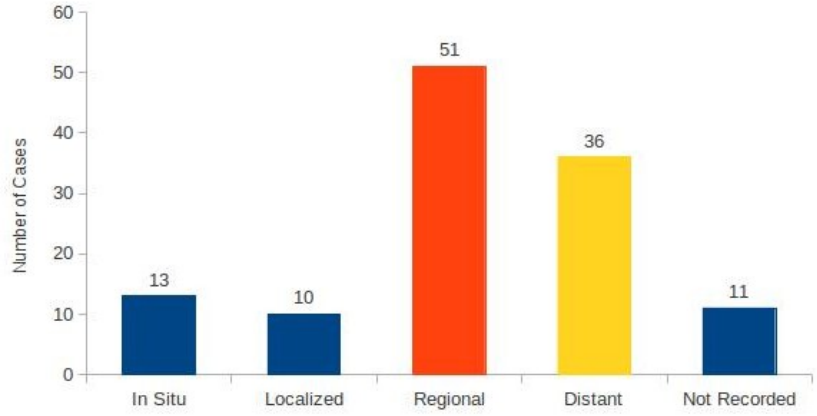


Figure 1.3: Distribution of oesaphageal cancer according to its extension[3].

They also reveal the lowest 5-year survival rate (6.7% and 5.5% respectively) when compared to the other extension groups. But the 1-year survival rate is drastically higher, 59.3% if

[4]This information was retrieved from http://seer.cancer.gov/tools/ssm/

regional, and 30.3% if metastatic[3].

It is thus crucial to detect oesophageal cancer, or any other type of cancer per say, in its early stages. In this context the NBI technique provides a valuable tool, emphasizing the mucosal microvasculature allowing the identification of vascular alterations indicative of a pathology. The use of two well defined wavelengths allows to visualize structures that were masked by conventional light. The possibility of observing structures lying deeper in the mucosa and the increase in contrast of superficial patterns that may be indicative of a pathology supply a very important visual tool for a correct early diagnosis. A CAD system capable of an accurate early detection is crucial in order to increase the odds for survival. Unfortunately, according to a study performed by Liedlgruber & Uhl (2011) the number of CAD related publications on the oesophagus is small when compared to other parts of the gastrointestinal tract. This number is even smaller when considering the new images acquired by NBI endoscopy.

1.2 Objectives

The increasing need of better decisions on the recognition of a pathology leads to the development of more accurate CAD system. In this context with this thesis we propose to study existing methods in the image processing field and apply them to NBI images. The special characteristics in the acquisition of NBI images leads to the need of developing new methods to describe these images in order to attain an improved classification performance. Rich information is therefore necessary and due to the fact that NBI is a quite recent technology, the lack of works done in this area is a big drawback.

In this thesis we study state of the art image processing techniques to describe gastroenterological images and develop a robust framework to classify these new images. Although, Computer Vision community usually use local or global analysis techniques for the processing and description of information, in gastroenterological images most research has been focused on global analysis techniques (a more detailed description will be given in Section 3.2). In this work we explore the usage of local image information.

The impact in classification of the independent information from the RGB channels is also studied by considering some physical principles in the image acquisition process. The

new patterns revealed by NBI and the high mortality of oesophageal cancer demands the development of a CAD system specialized in the determination of patterns that may be indicative of an evolution to cancer.

1.3 Thesis Outline

In Chapter 1 the NBI was introduced. The arising of NBI was natural in the sense of looking for ways of increasing the contrast between structures to allow a better early diagnostic. Some statistical data from RORENO is presented in order to motivate the need of the introduction of a CAD system appropriated to these new images.

In Chapter 2 we begin by introducing the physiological and anatomical properties of the oesophagus for a better understanding of the patterns found in the mucosa that will be characteristic of a possible pathology. All the images were in a pre-cancer stage and the used grading system is also presented. The scientific basis behind NBI is analysed and the main physical principles of the interaction of light with matter are introduced for a better understanding of the characteristics of the images and finally the Dichromatic Reflection Model are also presented.

Chapter 3 is dedicated to the process of extracting features from the images and we begin by presenting a literature review on this subject. In his thesis we propose to perform this task with local descriptors. We also present the sampling strategy that was used to extract rich information from the images.

Chapter 4 is dedicated to the presentation of some photometric invariants derived from the valid assumption of the Dichromatic Reflection Model. These photometric invariants are derived from the RGB colour model and we also derive the expressions for the opponent colour system as well as the respective invariants.

Chapter 5 is dedicated to the presentation of the learning method used in this thesis were we present some basic principles of the Support Vector Machine. In Chapter 6 we begin by presenting the separation of the images into two classes thus reducing to a binary classification problem. The main idea is to separate normal from abnormal cases. The used methodology to extract features, to build a vocabulary that will describe each image and the determination

of a proper set of descriptors to build our classifier, is also presented.

Next, in Chapter 7 we present the obtained results and we perform the discussion of the same and finally, in Chapter 8 we present the final conclusions of the developed work and future developments in the sense of improving the obtained results.

1.4 Contributions

The main contributions of the work presented in this thesis towards the recognition of pre-cancer lesions in gastroenterological images were the following:

1. Development of a framework for the representation of the images;
2. Analysis and assessment of the effectiveness of local descriptors;
3. Improvement of the recognition of pathologies through the addition of colour information based on physical models;
4. The developed work in this thesis was published at RecPad 2013 [5].

[5] http://soma.isr.ist.utl.pt/recpad/

Chapter 2

Background Knowledge

In this Chapter we begin to review the morphological and physiological characteristics of the oesophagus. In the first Section the structure of the oesophagus is described as well as the functionality of the different layers. It is also referred the main types of cancer found. Next, the visual patterns of the oesophagus that indicate a possible pathology are explained based on a simplified grading system of mucosal morphology against histology. We introduce the scientific basis behind NBI giving special focus on the absorption phenomenon that occurs due to the presence of haemoglobin in the capillaries. This absorption will have an impact on the formed image and thus it is critical to understand. We also review some basic physical principles mainly the ones that allow the understanding and influence the colours obtained from the image. Colour is in fact the crucial basis of this work and a more detailed description of the physical phenomenons that affect colour are referred as well as a physical model from whom all the images will be based. This model will be used in later chapters to build alternative colour spaces and derive photometric invariants.

2.1 The Oesophagus

The oesophagus is a flattened muscular tube of 18 to 26 cm in length. Microscopically, the oesophageal wall is composed of 4 layers: internal mucosa, submucosa, *muscularis propria* and *adventitia*. Unlike the remainder of the gastrointestinal tract, the oesophagus has no serosa. This fact allows tumours to spread more easily and make them harder to treat

surgically and also makes luminal disruptions more challenging to repair (Jobe et al., 2009).

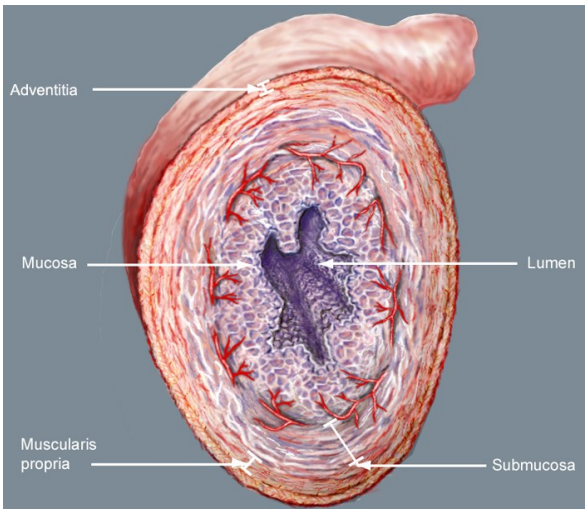


Figure 2.1: The layers in the oesophagus.

The internal mucosa is the most inner layer of the oesophagus and is considerably thick. It basically consists of the three sub-layers: the mucous membrane, which is a nonkeratinized squamous epithelium, *lamina propria* that contain vascular structures and the *muscularis* mucosa which is a thin layer of irregular arranged muscle fibres. As for the submucosa it mainly contains lymphocytes, plasma cells, nerve cells, a vascular network and submucosal glands. The *muscularis propria* lying more superficially then the submucosa is responsible for the motor function being composed exclusively of striated and smooth muscle. *Adventitia* is the most superficial layer of the oesophagus covering it and connecting it to neighbouring structures. It mainly contains small vessels, lymphatic channels and nerve fibres providing a support role (Jobe et al., 2009).

Oesophageal cancer, as referred in Section 1.1 is classified in two main groups: the Squamous Cell Carcinoma (SCC) and adenocarcinoma. These two types of cancer arise in different depths of the oesophagus. The first one, SCC occurs in the middle third of the oesophagus and the second one, adenocarcinoma is more common in the lower third of the oesophagus (Jobe et al., 2009). SCC seems to be the more frequent case of oesophageal cancer at least according to the data from RORENO representing 101 of the 121 diagnosed cases. The main visual signs that can be visualized with white light endoscopy or even better with NBI are the Gastroesophageal Reflux Disease (GERD) and Barrett's Oesophagus (BE). GERD is a symptom resulting from the upcoming of the gastric acid to the oesophagus. In its chronic

stage it is more likely to originate BE because repeated mucosal injury is thought to stimulate the progression of intestinal metaplasia. BE is defined as the replacement, or metaplasia, of the normal oesophageal squamous mucosa with a columnar epithelium containing goblet cells. It is the most important risk factor for oesophageal adenocarcinoma (Muto et al., 2009).

2.2 Grading System

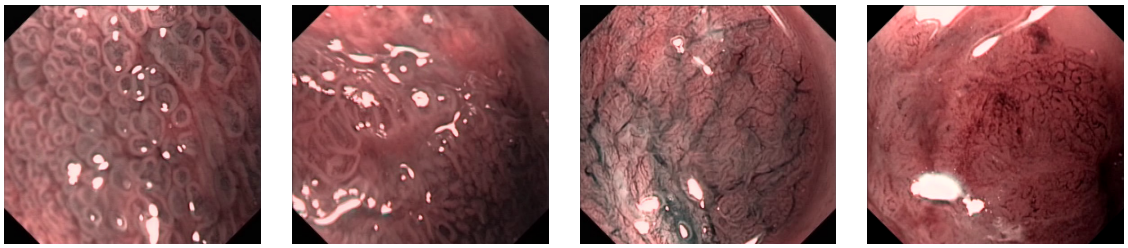
BE is one of the main indicators of a possible pathology in the oesophagus. In Singh et al. (2008) it was studied and validated a simplified grading system of the several patterns observed in BE. The system is based on the regularity of the patterns of the pits present in the mucosa as well as the patterns observed for the capillaries. The proposed system analyses images in a pre-cancer stage and classifies them into four distinct classes. As the mucosa starts to evolve into cancer it's surface becomes smoother, this is, the regular patterns start fading away. With their study they concluded that the patterns could be divided into four groups:

Type A : Round pits with regular microvasculature;

Type B : Villous/ridge pits with regular microvasculature;

Type C : Absent pits with regular microvasculature;

Type D : Distorted pits with irregular microvasculature;



Type A Type B Type C Type D

Figure 2.2: Sample images of our dataset following Singh's Grading System.

Type A images are normal and type B images are considered with low metaplasia. As for type C images the absent pits suggest a low dysplasia and type D are in a pre-cancer stage, high dysplasia. This simplified grading system was the base for this thesis to perform the separation of the images into the corresponding classes. They were previously classified by a cohort of experts and validated by histology.

2.3 The Theory of Narrow-Band Imaging

In the NBI technique the selection of two wavelengths that match the absorption peaks of haemoglobin will cause a maximum absorption of blue and green in different layers of the mucosa. Absorption occurs because part of the energy of the incident light is converted into heat through the vibrations of the molecules in the absorbing material. It is described by Lambert-Beer Law:

$$I(z) = I_0 \exp(-\mu_a z) \tag{2.1}$$

where $I(z)$ is the intensity of light after a path z along the tissue with μ_a absorption coefficient. I_0 is the incident intensity. One can define the absorption length L as the inverse of the absorption coefficient: $L = \frac{1}{\mu_a}$. This quantity measures the distance z in which the intensity $I(z)$ has dropped to $\frac{1}{e}$ of its incident value I_0. In Figure 1.1 we can see the absorption spectra of haemoglobin (HbO_2) which is predominant in vascularized tissues. We can observe four relative absorption peaks around 280 nm, 415 nm, 540 nm and 580 nm and a cut-off at approximately 600 nm (Niemz, 2007).

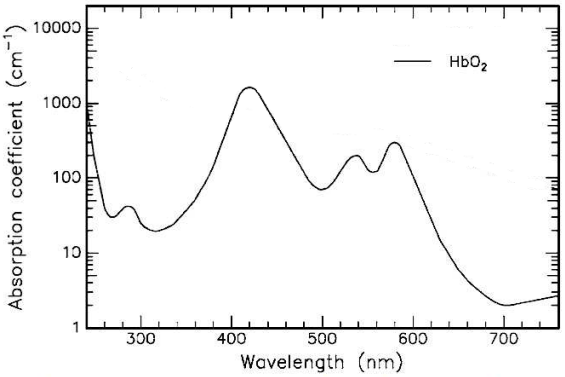


Figure 2.3: Absorption spectra of haemoglobin in blood (Figure from Niemz (2007)).

Note that the absorption length is wavelength dependent. Blue will be absorbed more superficially while green will be absorbed at a deeper layer. For this matter the resulting image can be thought as the combination of two independent absorption layers. The first one, more superficial corresponding to the blue light and a deeper one from the green light. This fact as a strong impact on the resulting image. The superficial capillaries will appear brown because the information of the blue is the input for the grenn and blue channels. The deeper vessels will appear cyan because the green is the input for both the red channel in the monitor. Figure 2.4 illustrates the different absorption lengths and the impact on the resulting image.

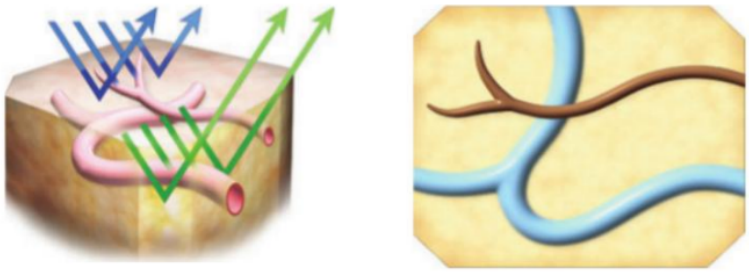


Figure 2.4: Absorption lengths in NBI and the enhancement of the capillaries (Figure from Bryan et al. (2008)).

While light penetrates in the tissue it interacts with small particles that may be cells, cell organelles and various fibre structures. Size, shape, density, their refractive index with respect to the tissue and the polarization states of these structures all interfere in the propagation of light in tissue. Scattering is the main factor that limits the imaging depth and contrast. Rayleigh scattering assumes that the distance between the scattering particles are much smaller than the wavelength of the incident radiation. The measured losses in light intensity due to scattering are quantified by an exponential decay law, defining a scattering coefficient μ_s. Neglecting the wavelength dependence of the index of refraction (thus neglecting dispersion phenomenons) we obtain Rayleigh's law of scattering:

$$I_s \sim \frac{1}{\lambda^4} \tag{2.2}$$

And taking the angle of scattering into consideration we obtain:

$$I_s(\theta) \sim \frac{1 + cos^2(\theta)}{\lambda^4} \tag{2.3}$$

If the spacing between the particles is comparable to the wavelength of the incident light, as is in blood cells (Niemz, 2007), another theory must be used: Mie scattering. The main difference to Rayleigh's scattering is the dependence on wavelength ($\sim \lambda^{-x}$ with $0.4 \leq x \leq 0.5$). But the probability of a photon to be scattered in a certain direction must be taken into consideration. For this matter Henyey–Greenstein proposed the following phase or probability function:

$$p(\theta) = \frac{1 - g^2}{(1 + g^2 - 2g\cos\theta)^{\frac{3}{2}}} \tag{2.4}$$

where g represents the coefficient of anisotropy and has the values 1 or -1, forward and backward scattering respectively. If $g = 0$ isotropic scattering occurs. Typical values for g range from 0.7 to 0.99 for biological tissues and the corresponding scattering angles are 8^o and 45^o (Niemz, 2007).

Most biological tissues are turbid and therefore both scattering and absorption will occur. The mucosa is no exception. For such media one must then define a total attenuation coefficient as:

$$\mu_t = \mu_a + \mu_s \tag{2.5}$$

thus considering the contributions of scattering (μ_s) and absorption (μ_a). One can also define the mean free optical path of the photons through the mucosa as:

$$L_t = \frac{1}{\mu_t} \tag{2.6}$$

In order to have a better idea if a medium is mostly absorbing or scattering, and thus the attenuation of light is mostly due to absorption or scattering, it is usually defined another parameter, the optical albedo a, given by:

$$a = \frac{\mu_s}{\mu_t} \tag{2.7}$$

If $a = 0$, attenuation is mostly due to absorption, if $a = 1$ attenuation is mostly due to scattering and if $a = \frac{1}{2}$ both occur.

In the literature it is common to work with the reduced scattering coefficient, defined as:

$$\mu_s^{'} = \mu_s(1 - g) \tag{2.8}$$

Another useful parameter is the optical penetration depth, defined as:

$$d = \int_0^s \mu_t ds^{'} \tag{2.9}$$

where ds' is an infinitesimal segment of the optical path and s is the total length (Niemz, 2007). These definitions are very useful for an experimental determination, using for example, the inverse adding-doubling method. This was performed by Bashkatov et al. (2005). They determined some optical properties of human skin, subcutaneous adipose tissue and human mucosa for a wavelength window of 400 nm to 2000 nm. Borrowing the data related to the human mucosa from their work we observe that the absorption coefficient of the human mucosa presents two peaks at approximately 415 nm and 540 nm due to the presence of haemoglobin in the oxygenated form in the superficial vessel of the mucosa. The reduced scattering spectra actually reveals that for the used wavelengths the mucosa is a quite scattering tissue and presents an anomalous behaviour near the absorption peaks. This is due to an anomalous light dispersion phenomenon. It is also observed that the penetration depth of light at the referred wavelengths is very superficial.

If we illuminate the tissue with white light some of the vascular structures are not visible neither are other patterns that maybe indicative of a tissue lesion. But narrowing the band of the light output will reduce scattering effects and increase the image definition and with the absorption phenomenon the contrast between superficial and deeper vessels is achieved providing a high quality tool for a better diagnosis.

2.4 The Dichromatic Reflection Model

When a light ray hits a surface of a material part of it will be reflected and the remaining will penetrate throw the tissue. While light penetrates in the tissue it can be scattered, absorbed or in a more realistic case, a little bit of both. If we consider the angle of incidence as being θ_i and the angle of reflection as θ_r with respect to the normal of the incidence plane, we have $\theta_i = \theta_r$. This is the first part of the Law of Reflection. The second part states that the incident ray, the perpendicular of the surface and reflected ray all lie in a plane called the plane of incidence. Ignoring polarization we can obtain the reflectivity or reflectance given by for normal incidence:

$$R = r^2 = \left(\frac{n_0 - n_1}{n_0 + n_1}\right)^2 \tag{2.10}$$

where n_0 and n_1 are the refractive indices of the incidence and transmission media and they are wavelength dependent, so R will vary along the spectrum. If a reflecting surface

is smooth, that is, the irregularities, are small compared to the wavelength, the light re-emitted by the millions of atoms will combine to form a well defined beam in a process called specular reflection. On the other hand if the surface is rough compared to the wavelength the emerging rays will have different directions constituting what is called diffuse reflection (Hecht, 1998).

The Dichromatic Reflection Model (DRM) was originally proposed by Shafer (1985). Before introducing the model it is important to make some considerations. The tissue constituting the surface of the mucosa can be modelled considering it to be optically inhomogeneous. This means that the light will interact with the surface matter and with particles of a called colourant, responsible for producing scattering and colouration. The surface is also considered opaque, this is, it does not transmit light from one side to the other.

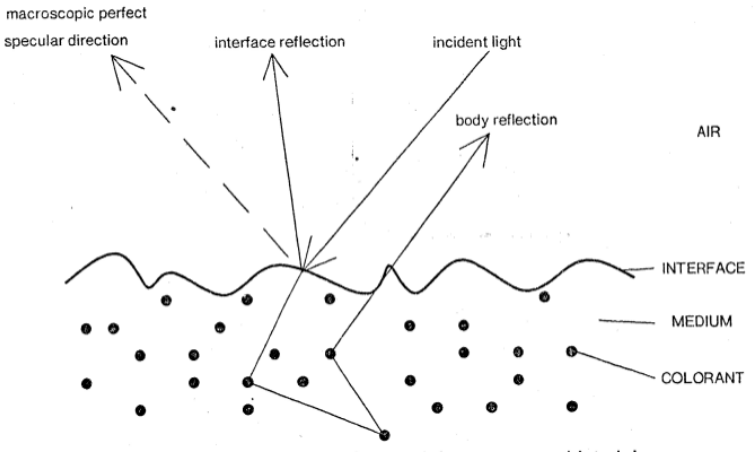


Figure 2.5: Reflection of light from an inhomogeneous material (Figure from Shafer (1985)).

As light hits the interface of the mucosa it encounters a different index of refraction and therefore some of it will be reflected according to Equation 2.10. In a first approximation the reflection will be in the perfect specular direction. The surface is optically rough, meaning that the local surface normals are different from the considered surface normal and hence light will be slightly scattered. The reflection that occurs at the surface is called *interface reflection*. Interface reflection is said to be constant with respect to wavelength and thus has the same colour as the illuminant (Shafer, 1985).

While light penetrates the mucosa it can be scattered by the colourant, and be transmitted through the mucosa, it can be absorbed by haemoglobin molecules or it can be re-emitted

through the mucosal surface, producing the *body reflection* as illustrated in Figure 2.5. The geometric distribution of light resulting from the body reflection is considered isotropic, meaning independent of the viewing angle. Its colour will be different from the illuminant because absorption or scattering occurs and their probabilities are wavelength dependent. The DRM then considers that the image is formed by two terms: interface and body reflection. The illumination direction and surface normal are represented by I and N. V is the viewing direction and J is the direction of the macroscopic perfect specular reflection. The photometric angles are i, the incidence angle, e is the angle of emittance between N and V, g is the phase angle between I and V and s is the off-specular angle defined between the viewing direction N and the direction of the perfect specular reflection J. The reflectance geometry is shown in Figure 2.6.

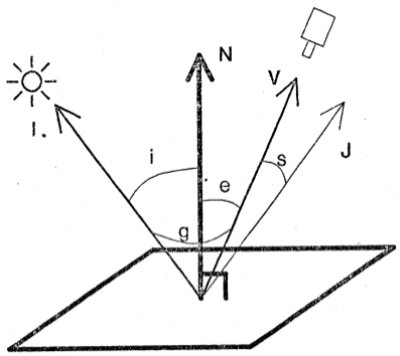


Figure 2.6: Photometric Angles (Figure from Shafer (1985)).

The model states that:

$$L(\lambda, i, e, g) = L_i(\lambda, i, e, g) + L_b(\lambda, i, e, g) \tag{2.11}$$

where L is the total radiance of the reflected light, L_i is the interface reflection term and L_b is the body reflection term. Both radiances can be decomposed into a relative spectral power distribution (*composition*) c_i and c_b, that depend only on the wavelength and are independent of geometry, and a geometric scale factor (*magnitude*) m_i and m_b that depend only on geometry. Equation 2.11 can then be rewritten as:

$$L(\lambda, i, e, g) = m_i(i, e, g)c_i(\lambda) + m_b(i, e, g)c_b(\lambda); \tag{2.12}$$

The validity of the model assumes certain circumstances such as the inhomogeneity and opacity of the surface, a uniform distribution of the colourant, no fluorescence or thin-film

properties, isotropic reflection from the surface, no inter-reflection among surfaces, no diffuse illumination and a relative spectral power distribution of the illumination across the scene. The generality of the model is based on not making assumptions on the geometry imaging, the fact that it doesn't specify if the surface is planar or curve, it does not assume any specific functions for the terms m_i, m_b, c_i and c_b. It also applies equally well if we have a point light source, an extended light source or infinitely far way light source. Finally it does not assume that the amount of illumination is the same everywhere in the scene but only the spectral power distribution to be the same thus approximating to a more realistic situation (Shafer, 1985).

When a sensing device such as a camera records an image, light is integrated over the entire spectrum. This process of spectral integration sums the amount of incoming light $L(\lambda, i, e, g)$, weighted by the spectral transmittance of the filter, $\tau(\lambda)$, and the spectral responsivity of the camera, $s(\lambda)$, over all the wavelengths:

$$C = \int_{\lambda} L(\lambda, i, e, g)\tau(\lambda)s(\lambda)d\lambda \tag{2.13}$$

By using a red, green and blue filters we are reducing the problem to a three dimensional vector space. The spectrum of an incoming beam at pixel position (x, y) is represented by a triplet $C(x, y) = [R, G, B]$ where i, e and g are determined by x and y and by the position of the object relative to the camera. Spectral integration is a linear transformation (Shafer, 1985). For this matter Equation 2.12 still holds after the spectral integration and we can write the DRM in a three dimensional space as follows:

$$C(x, y) = m_i(i, e, g)C_i + m_b(i, e, g)C_b \tag{2.14}$$

$C(x, y)$ is thus a linear combination of two three dimensional vectors, $C_i = [R_i, G_i, B_i]$ and $C_b = [R_b, G_b, B_b]$, that spans the dichromatic plane in the three dimensional colour space. Within the defined parallelogram by C_i and C_b, the position of any colour is defined by the magnitude of the coefficients m_i and m_b at the corresponding point. The DRM thus provides a mathematical tool to predict the distribution of colours in an object.

Chapter 3

Invariant Features

In this chapter we begin to review some of the literature related to techniques to extract features from the images. These techniques can be grouped according to their domain: spatial or frequency, or according to the level they are extracted: high or low level features. For this task we propose the usage of a local descriptor based on spatial domain features. These features are proved to be invariant to scale, rotation and translation transformations (Lowe, 2004).

3.1 Related Work

In this section we review some of the work done in the extraction of features and the different methods that several authors have used in their works to treat and classify all the information contained in an image, definitely worth more then a thousand words.

Spatial Domain Features: In Sousa et al. (2009) adapted colour features combined with local binary patterns were used in order to build a texture descriptor to natural endoscopic images. This work was based on Dinis-Ribeiro visual classification for gastric mucosa and used statistical pattern recognition methodologies to mimic this visual work done by clinicians. An MPEG-7 visual descriptor for feature extraction in capsule endoscopy was analysed by Coimbra & Cunha (2006). MPEG-7 is a multimedia content description standard

that defines a variety of visual descriptors for video classification that mainly divide in two groups, colour descriptors and texture descriptors. They showed that the Scalable Colour and Homogeneous Texture descriptors are the most adequate to visually detect an event in capsule endoscopy videos. In Poh et al. (2010) a fusion of low-level features with intermediate-level features was presented. Their work was to classify Wireless Capsule Endoscopy (WCE) images as bleeding or non-bleeding. At a low-level they divided the image into square sized cells of pixels characterizing it with an adaptive colour histogram and then using a cell-classifier. In an intermediate-level they divided the image in blocks and then classified each block combining this information with the previous one. The conclusions were that this multi-level system actually improved the information representation for WCE images.

Frequency Domain Features: Features are extracted from an image or from the colour channel after applying some transformation to the data in the frequency domain. Colour wavelet features were used by Karkanis et al. (2003) to extract information from endoscopic video images. The features were based on covariances of second-order textural measurements calculated over the wavelet frame transformation of different colour bands. Expanding this work, in Lima et al. (2008) a third and forth order moments were added to cope with distributions that tend to become non-Gaussian in some pathological cases. They achieved 95% specificity and 93% sensitivity although only 6 full endoscopic exams were used. Also in this context, in Kwitt & Uhl (2007) a feature extraction method based on fitting a two parameter Weibull distribution to the wavelet coefficient magnitudes of sub-bands was presented. They assumed textural measures from zoom-endoscopy images that were calculated from the sub-bands of a complex wavelet transform variant known as the Dual-Tree Complex Wavelet Transform. They claimed a significant improvement of the leave-one-out cross-validation (LOOCV) accuracy compared with the classic mean and standard deviation features. Texture has also been the main property evaluated by Karkanis et al. (2001). They used second order statistics of the wavelet transformation of each video-frame. That information was estimated utilizing co-occurrence matrices that were textural signatures of the corresponding regions. A big difference to other works is that they used a multi layer feed forward Neural Network (MFNN) architecture which was trained using features on normal and tumour regions. Another approach was proposed by Khademi & Krishnan (2007). They extracted statistical features from the wavelet domain describing the

homogeneity of areas in small bowel images. They explored a shift-invariant discrete wavelet transform (SWIDWT) claiming a high classification rate.

High-level Features: Other authors prefer to work with features that are not extracted based on colour or texture properties but instead they describe geometrical properties of shapes extracted from the images. They usually use an edge detector algorithm like the Canny or SUSAN edge detectors. Some authors have also used a different approach with region-based algorithms such as segmentation or region growing. An extension of a state-of-the-art algorithm for boundary detection and segmentation in application to colonoscopic NBI images that perform automatic segmentation to the images was proposed by Ganz et al. (2012). Another approach was done by Coimbra et al. (2010) were they measured the impact of several segmentation algorithms and performing an automatic classification of gastric tissue. Another one was done by Karargyris & Bourbakis (2009) using Log Gabor filters in Wireless Capsule Endoscopy videos. The idea was to use a SUSAN edge detector in conjunction with a Log Gabor filter to automatically detect the presence of polyps. In Stehle et al. (2009) the extraction of vascularization features using and comparing different segmentation algorithms to vessels were performed concluding that the phase symmetry and the fast marching algorithms gave the best results.

3.2 Scale Invariant Feature Transform

In this thesis we propose a different approach based on the spatial domain features. For our local descriptor we used Scale Invariant Feature Transform (SIFT). It was first introduced by Lowe (1999) in the attempt of providing a method to extract features that could be used for image matching. The extracted features showed to be invariant to scale and rotation and provided a robust image matching in a wide variety of transformations in the image due to the distinctiveness of the features (Lowe, 2004).

The SIFT approach can be divided into two parts. The first one consists in detecting points of interest or key-points by smoothing the image with different levels and making the difference in order to detect a maxima or a minima, in a process called SIFT detector. This approach intends to mimic the object recognition process in primates based on invariant features to

scale, rotation and illumination. Some neurons in the inferior temporal cortex highly respond to shape features that share the same complexity of the SIFT features (Lowe, 1999).

The second part is the description of the image around the detected key-point achieved by stacking spatial gradient histograms descriptive of an image region. This part is called the SIFT descriptor.

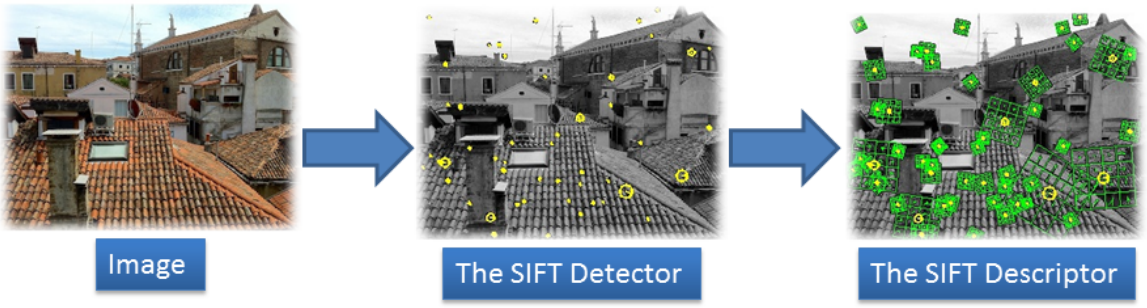


Figure 3.1: The SIFT method (Adopted from (Vedaldi & Fulkerson, 2008).

3.2.1 The SIFT detector

The first step in detecting the key-points is to find locations that are invariant to scale changes. This is done by searching for stable features across different scales of the image using a continuous function of scale know as scale space (Lowe, 2004). It is also proved that the only possible function that assures this invariance is the Gaussian function. The scale-space of an image can then be built by convoluting the image with a Gaussian function. Suppose our image is defined as $I(x, y)$ and the scale dependent Gaussian function as $G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{\frac{-(x^2+y^2)}{2\sigma^2}}$. The scale space of an image, $L(x, y, \sigma)$ is defined as:

$$L(x, y, \sigma) = G(x, y, \sigma) * I(y, y) \tag{3.1}$$

By smoothing the image with a Gaussian function, (Lowe, 1999) intends to obtain similar results with the response of some neurons to colour and texture.

The idea is to compare nearby scale-spaces of an image in order to assure the stability of the key-point. This is done using the Difference-of-Gaussian (DoG) function, $D(x, y, \sigma)$ convolved with the image. By taking two nearby scales separated by a multiplicative factor k, the DoG is simply the subtraction of the images convolved by the two scales respectively,

and is given by:

$$D(x, y, \sigma) = L(x, y, k\sigma) - L(x, y, \sigma) \tag{3.2}$$

Each octave of the scale space is divided into an integer number, s, so that $k = 2^{\frac{1}{s}}$. To cover a whole octave we need $s + 3$ scale-space images and two adjacent ones are then subtracted to compute the DoG images. Once a complete octave is processed, the scale-space images are down-sampled by two and the process is repeated. This means that increasing the scale by an octave is the same as halving the image resolution (Vedaldi & Fulkerson, 2008). Figure 3.2 illustrates this methodology.

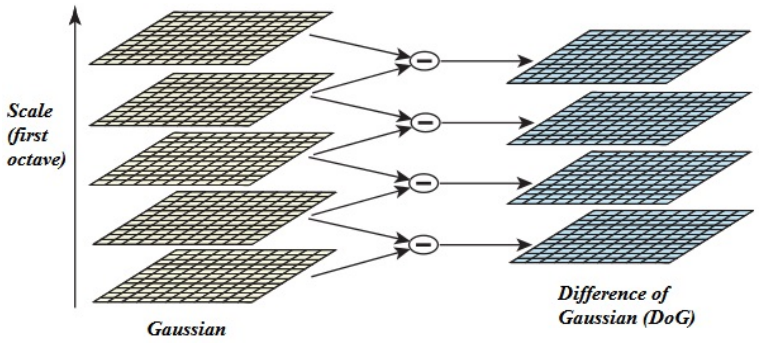


Figure 3.2: The computation of the DoG. (Figure from Lowe (2004)).

Maxima and minima of the DoG are determined by comparing a sample point to it's 8 nearest neighbours in 3x3 regions at the current image and nine neighbours in the scale above and below. If this point is the greatest or the smallest of all them, then it's a candidate for a key-point.

To achieve the desired stability of the key-points not all of them will be considered. Points that present low contrast, thus sensitive to noise, are rejected by using a Taylor expansion up to second order of the DoG function. Points that are poorly localized in an edge are also discarded by taking the curvature of the peak computed with a 2×2 Hessian matrix, at the location and scale of the key-point. A bad key-point will have a large curvature across the edge but a small one at the perpendicular direction (Lowe, 2004).

Once the key-point is determined, the magnitude of the gradient, $m(x, y)$, and its orientation, $\theta(x, y)$ are computed in order to achieve invariance to rotation and to scale, and they are computed using pixel differences by the following expressions (Lowe, 2004):

$$m(x, y) = \sqrt{((L(x + 1, y) - L(x - 1, y))^2 + (L(x, y + 1) - L(x, y - 1))^2} \tag{3.3}$$

$$\theta(x,y) = tan^{-1}\left[\frac{L(x,y+1) - L(x,y-1)}{L(x+1,y) - L(x-1,y)}\right] \tag{3.4}$$

The SIFT detector is a circle of radius equal to the scale. A geometric frame of four parameters will describe the key-point. The x and y position of the center of the key-point determined by Equation 3.3, it's scale, s and orientation θ given by Equation 3.4.

3.2.2 The SIFT descriptor

The determined scale will influence the level of the Gaussian blur at which the image is smoothed and the orientation of the key-point will rotate the descriptor geometry. Around the key-point a sample array of 16×16 is defined and for each one of this regions the gradient magnitude and orientation is computed according to Equations 3.3. The magnitude of the gradients are then stacked into 8 possible orientations thus summarizing the information of a 4×4 region as shown in Figure 3.3.

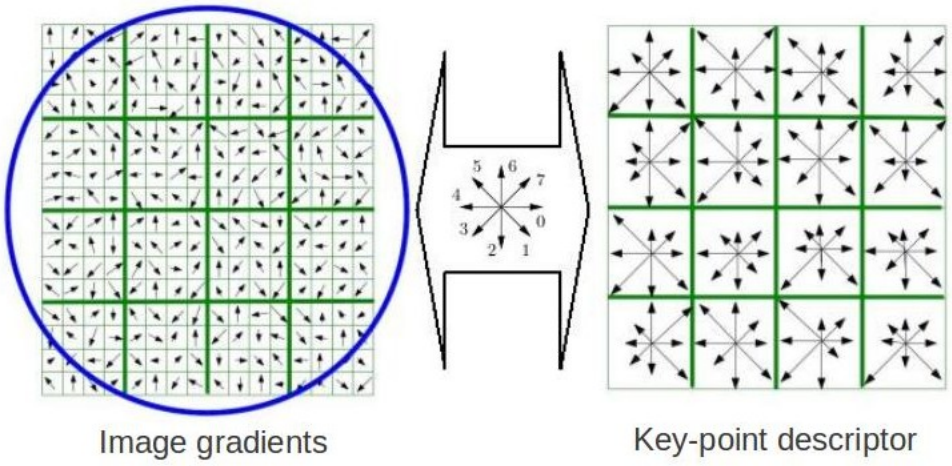


Figure 3.3: The SIFT descriptor.

The length of the arrows shown in Figure 3.2 on the right are the result of the summation of the magnitudes of the gradients to their closest orientations according to the 8 bin orientation histogram in the middle. A Gaussian weighting function is also applied to give less importance to the gradients farther away from the center of the key-points. Each descriptor around the key-point will thus have $8 \times 4 \times 4 = 128$ dimensions. This way of sampling the image gradients into 4×4 regions is proved by Lowe (2004) to give the best results.

3.2.3 Sampling Strategies

The SIFT approach consists in calculating the position of a key-point by computing the DoG and extract the desired features around that key-point. The features consist of the computed gradients stacked in a spatial histogram that contains information of a 4x4 region. But a question arises. What is the best sampling strategy to use in this set of images? Should we determine the location of the key-point or could we skip that part of the SIFT approach and perform a different type of sampling? According to Nowak et al. (2006) the performance of a classifier based on SIFT features increases with an increasing number of points per image. They also concluded that for a high number of points the best way of sampling an image was to perform a uniform random sampling (Nowak et al., 2006). With this in mind and knowing we have images whose textures are spread over a region, thus having a considerable number of points per image, we chose to perform a dense sampling of the SIFT descriptors. This way of densely sampling an image is called a DSIFT. It consists in skipping the detection stage of the key-points. The idea is to place a quadrangular grid of key-points on top of the image and extract SIFT descriptors around each key-point. The regularly placed key-points are all at the same scale and orientation and so are the extracted descriptors.

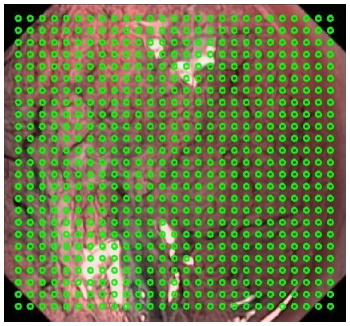


Figure 3.4: Dense sampling strategy.

Figure 3.4 is an example of applying DSIFT to an image. We see the quadrangular grid of key-points placed on top of the image. The extracted SIFT descriptors will be placed at the center of each key-point. The radius of the circles is the scale of the key-point which in turn is the parameter for the smoothing Gaussian. The orientation of the key-point which in this type of sampling is equal to every key-point will be also the orientation of the extracted descriptor[2].

[2]More technical details will be explained further in Chapter 6.

Chapter 4

Adding Colour Information

With the Dichromatic Reflection Model, presented in Section 2.4, in this Chapter we seek alternative formulations for the RGB colours in an attempt to improve the results obtained by a regular conversion to gray-scale. The RGB colour system and some of its properties are briefly presented as well as the opponent process theory and the derivation of the mathematical expressions of the opponent colours. These colour systems are endowed by important photometric invariance properties that augment the robustness of posterior recognition process.

4.1 Colour Spaces

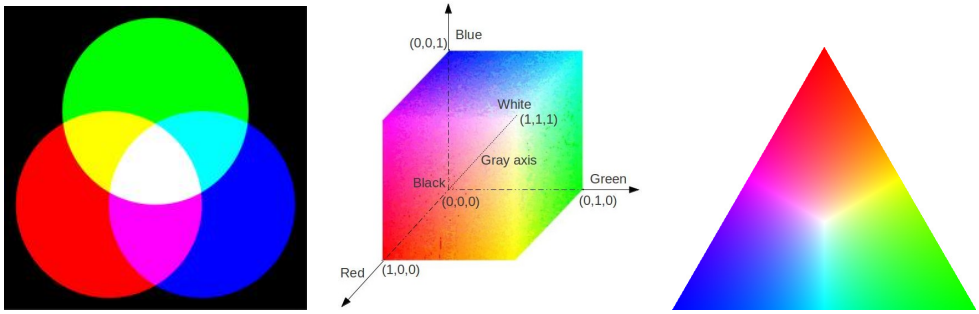
We saw in Section 2.4 that the spectrum of an incoming beam at position (x, y) is given by Equation 2.13. Most of the device manufacturers opted to use a red, green and blue filter and the sensitivities of each filter are reasonably well matched to the human eye. By doing so, we are reducing the problem to three dimensions and thus the spectrum of the incoming beam is represented by a triplet $C = [R, G, B]$.

RGB Colour System: The RGB colour system is an additive colour model. A broad array of colours can be reproduced by adding certain amounts of red, green and blue. We can then build a colour cube defined by the R, G and B axis. White is produced when all colours are

at a maximum light intensity. Black on the other hand is produced when all colours are at a minimum light intensity, the origin of the referential. The diagonal connecting the white and black corners of the RGB cube defines the intensity:

$$I(R,G,B) = R + G + B \tag{4.1}$$

Figure 4.1 (a) shows the addictiveness of the RGB colours and Figure 4.1 (b) shows the RGB colour cube.



(a) (b) (c)

Figure 4.1: (a) The addictiveness of the RGB system. (b) The RGB cube. (c) The chromaticity triangle.

All the points in a plane perpendicular to the gray axis have the same intensity. Such plane cuts an equilateral triangle defining the rgb chromaticity triangle (Figure 4.1 (c)) given by:

$$\begin{pmatrix} r(R,G,B) \\ g(R,G,B) \\ b(R,G,B) \end{pmatrix} = \begin{pmatrix} \frac{R}{I} \\ \frac{G}{I} \\ \frac{B}{I} \end{pmatrix} \tag{4.2}$$

This normalization allows one to calculate colour features from the original R, G and B values from the corresponding red, green and blue images provided by the colour camera. Briefly, on Section 4.2, we will see the advantages of using the rgb normalized colour system.

Hue, Saturation and Intensity: One of the definitions of hue is that it is described by the dominant wavelength of a spectral power distribution. Saturation is usually defined as the purity of a colour, decreasing when more achromaticity is mixed into a colour. Completely desaturated colours coincide with the gray axis, while fully saturated colours coincide with pure colours. Intensity is defined, for the RGB colour system by Equation 4.1.

Opponent Colour System: Human beings do not rely on long (L), middle (M) and short (S) wavelengths channels like the RGB system assumes. The human retina possesses ganglion cells that combine L, M and S channels to work in opponent colours in order to enhance the detection of events of interest. The opponent colour theory started about the year 1500 when Leonardo da Vinci came to the conclusion that colours were produced by a mixture of red-green, yellow-blue and white-black. This opponent colour theory was completed by Edwald Hering in late 19th century (Gevers et al., 2012). He stated that humans had three types of photo receptors corresponding to each pair of the opponent colours. But nowadays, we now that those cells do not exist, but he was right about one thing, the computation of opponent colours (Gevers et al., 2012).

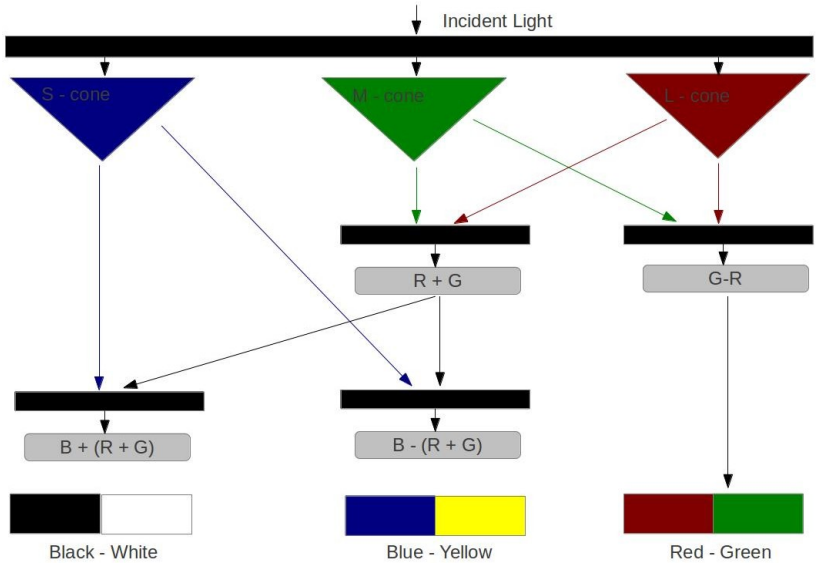


Figure 4.2: The opponent process theory.

The colour photo receptors at the retina, the cones, are sensitive to long, L-cone, middle, M-cone and short, S-cone, wavelengths. A single cone is colour blind since its activation depends on both the wavelength and the intensity of the stimulus. A comparison of the signals from different classes of photo receptors is therefore the basis for a vision system. At an early stage in the red-green opponent pathway, signals from the L and M cones are opposed and in the yellow-blue pathway signals from S cones oppose a combined signal from L and M cones. In addition, there is a third luminance or achromatic mechanism in which retinal ganglion cells receive L and M cones input. Thus L, M and S belong to a first layer of the retina whereas luminance and opponent colours belong to a second layer, forming the

basis of chromatic input to the visual primary cortex.

Despite this physiological interpretation of the opponent colours we can get a more mathematical derivation based on photometric derivatives. By considering the DRM referred in Section 2.4 we can take the spatial derivative of the expression given by Equation 2.14 to get:

$$C^{'} = em_b c_b^{'} + (e^{'} m_b + em_b^{'})c_b + (em_i^{'} + e^{'} m_i)c_i \tag{4.3}$$

This photometric derivative can be seen as a combination of three vectors. Lets take a closer look at each one of them (Figure 4.3).

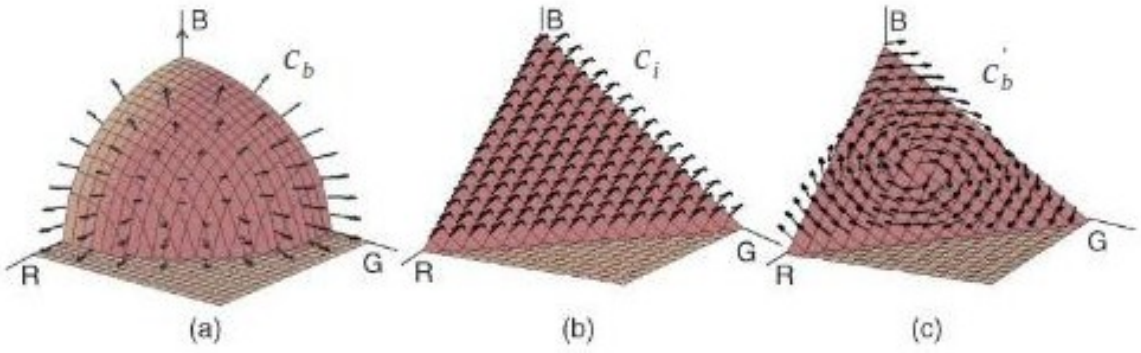


Figure 4.3: (a) Shadow-shading direction. (b) Specular direction. (c) Hue direction. (Adopted from Van De Weijer et al. (2005)).

First lets look at the term dependent on c_b, the second term. What we have are two different physical phenomenons. For example, if we have changes in the illumination direction and a fixed object, this is we have $e^{'} m_b$, we will have a shadow. On the other hand if we have a fixed illumination direction but a varying geometry coefficient we have a shading. This vector direction at which this changes occur is then called the shadow-shading direction. In the specular direction, where changes in m_i occur, we also have two terms, one corresponding to changes of the illumination direction, $e^{'} m_i$ which represents a shadow on top of a specular reflection and another that corresponds to changes in the geometric coefficient $em_i^{'}$. We can also define a third direction that will be perpendicular to this last two and that is where the hue direction arises, the direction of the vector $c_b^{'}$ (Van De Weijer et al., 2005).

If we transform the RGB colour space coordinates by means of a spherical transformation and taking the shadow-shading as one if its coordinates, we get the $r\theta\phi$ colour space given

by:

$$\begin{pmatrix} r \\ \theta \\ \phi \end{pmatrix} = \begin{pmatrix} \sqrt{R^2+G^2+B^2} = |C| \\ \arctan \frac{G}{R} \\ \arcsin \frac{\sqrt{R^2+G^2}}{\sqrt{R^2+G^2+B^2}} \end{pmatrix} \tag{4.4}$$

On the other hand if we take the specular direction as a component of a new orthogonal space we get the opponent colour space, which for a known illuminant $c_s = (\alpha, \beta, \gamma)^T$, is given by:

$$\begin{pmatrix} O_1 \\ O_2 \\ O_3 \end{pmatrix} = \begin{pmatrix} \frac{\beta R - \alpha G}{\sqrt{\alpha^2+\beta^2}} \\ \frac{\alpha\gamma R + \beta\gamma G - (\alpha^2+\beta^2)B}{\sqrt{(\alpha^2+\beta^2+\gamma^2)(\alpha^2+\beta^2)}} \\ \frac{\alpha R + \beta G + \gamma B}{\sqrt{\alpha^2+\beta^2+\gamma^2}} \end{pmatrix} \tag{4.5}$$

By taking white illumination the source term is $c_s = (1, 1, 1)^T$, and so the opponent colour formulas simplify into:

$$\begin{pmatrix} O_1 \\ O_2 \\ O_3 \end{pmatrix} = \begin{pmatrix} \frac{R-G}{\sqrt{2}} \\ \frac{R+G-2B}{\sqrt{6}} \\ \frac{R+G+B}{\sqrt{3}} \end{pmatrix} \tag{4.6}$$

If we take the hue direction as a component for a new coordinate system we get the hue-saturation-intensity for the opponent colours, simply by considering a polar coordinate transformation of the Red Green Blue (RGB) colour space and are given by:

$$\begin{pmatrix} h \\ s \\ i \end{pmatrix} = \begin{pmatrix} \arctan\left(\frac{O_1}{O_2}\right) \\ \sqrt{O_1^2+O_2^2} \\ O_3 \end{pmatrix} = \begin{pmatrix} \arctan\left(\frac{\sqrt{3}(R-G)}{R+G-2B}\right) \\ \sqrt{\frac{4}{6}(R^2+G^2+B^2-RG-RB-GB)} \\ \frac{R+G+B}{3\sqrt{3}} \end{pmatrix} \tag{4.7}$$

O_1 roughly corresponds to the red-green channel, O_2 to the yellow-blue channel and O_3 to the intensity channel. The opponent colour system largely decorrelates the RGB colour channels although it is device dependent and it is not perceptually uniform, this is, the numerical distance between to colours cannot be related to perceptual differences.

4.2 Photometric Invariant Features

The DRM can be applied to derive photometrically invariant features. If we assume only matte, or dull surfaces, specular reflection is negligible, this is $m_i = 0$. The angles i, e and

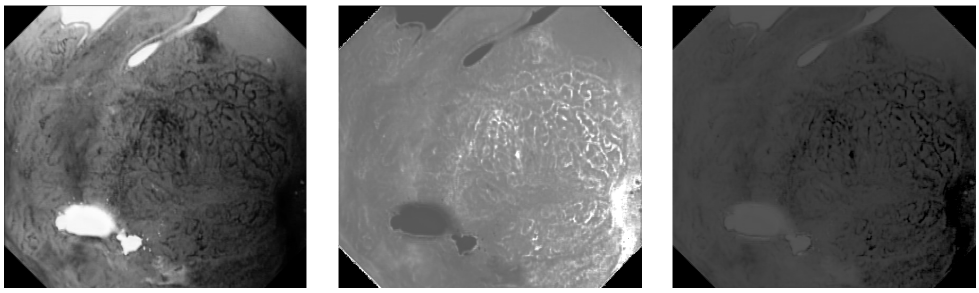
g fully specify the location of the pixel so we can simplify the notation by denoting ρ as the spatial coordinates, and so Equation 2.14 reduces to the Lambertian model for diffuse body reflection:

$$C(\rho) = m_b(\rho)C_b \tag{4.8}$$

A zero-order invariant can be obtained building each channel with the assumption given in Equation 4.8. This way the normalized rgb can be considered invariant to lighting geometry and viewpoint, this is, independent of m_b, since:

$$r = \frac{R}{R+G+B} = \frac{m_b(\rho)C_b^R}{m_b(\rho)(C_b^R + C_b^G + C_b^B} = \frac{C_b^R}{C_b^R + C_b^G + C_b^B} \tag{4.9}$$

Similar equations can be obtained for the normalized g and b. It thus results in the independence for the surface orientation, illumination direction and illumination intensity, assuming Lambertian reflection and white illumination. This normalized rgb space depends only on the factors C_b^R, C_b^G and C_b^B which depend on the sensor and the surface albedo (Gevers et al., 2012).



gray r g

Figure 4.4: Gray-scale image and the r and g channels.

In fact with the DRM applied to each channel we can perform any linear combination. One of the reasons to do so is the possibility to capture intensity variations in regular surfaces. This way we have to analyse the proportion of these variations in order to diminish this dependency. That said we can compute:

$$\begin{aligned} C_{RGB} &= \frac{\sum_i a_i (C^R)_i^p (C^G)_i^q (C^B)_i^r}{\sum_j b_j (C^R)_j^s (C^G)_j^t (C^B)_j^u} \\ &= \frac{\sum_i a_i (m_b(\rho)C_b^R)_i^p (m_b(\rho)C_b^G)_i^q (m_b(\rho)C_b^B)_i^r}{\sum_j b_j (m_b(rho)C_b^R)_j^s (m_b(\rho)C_b^G)_j^t (m_b(\rho)C_b^B)_j^u} \\ &= \frac{\sum_i a_i m_b(\rho)^{p+q+r} (C^R)_i^p (C^G)_i^q (C^B)_i^r}{\sum_j b_j m_b(\rho)^{s+t+u} (C^R)_j^s (C^G)_j^t (C^B)_j^u} \end{aligned} \tag{4.10}$$

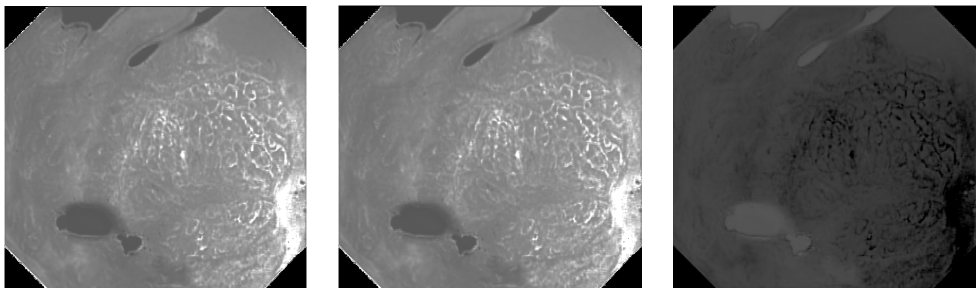
Since $p+q+r=s+t+u$, Equation 4.10 can be further simplified to:

$$C_{RGB} = \frac{\sum_i a_i (C_b^R)_i^p (C_b^G)_i^q (C_b^B)_i^r}{\sum_i b_i (C_b^R)_j^s (C_b^G)_j^t (C_b^B)_j^u} \tag{4.11}$$

Numerous invariants can be obtained. The set of first-order invariants involves the set where $p+q+r=s+t+u=1$:

$$\left\{ \frac{R}{G}, \frac{R}{B}, \frac{G}{B}, \cdots \right\} \tag{4.12}$$

This set of invariants show the same invariance properties as the normalized rgb colours.

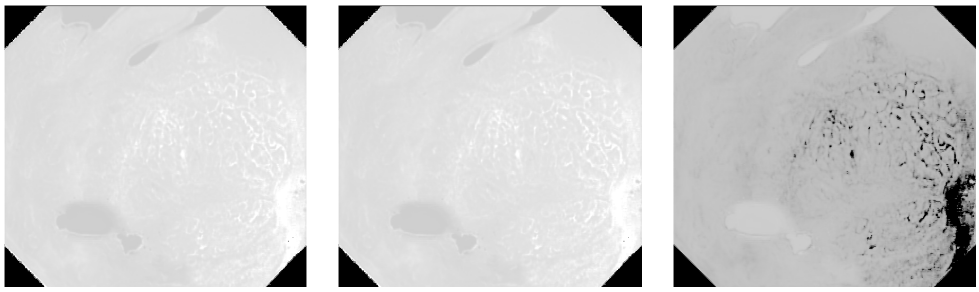


$\frac{R}{G}$ $\frac{R}{B}$ $\frac{G}{B}$

Figure 4.5: The ratio of the channels.

It happens that there are quick density variations that cannot be captured linearly. So in this work we introduce the proportion under the logarithm for this end.

$$\left\{ \log\left(\frac{R}{G}\right), \log\left(\frac{R}{B}\right), \log\left(\frac{G}{B}\right), \cdots \right\} \tag{4.13}$$



$\log\left(\frac{R}{G}\right)$ $\log\left(\frac{R}{B}\right)$ $\log\left(\frac{G}{B}\right)$

Figure 4.6: The logarithm of the ratio of the channels.

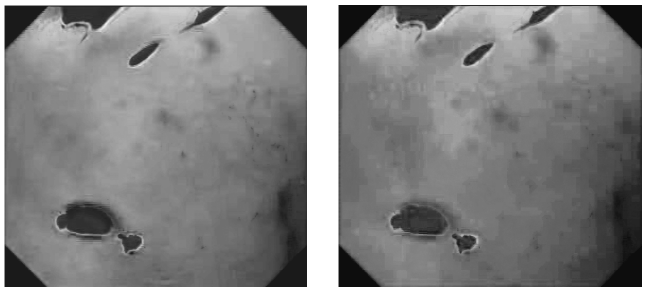
Regions with low illumination or irregular surfaces not clearly illuminated may have lesions and hence gradient information cannot be explicitly extracted. Performing a non-linear

mapping on the combined colour system will provide the necessary enhancement to properly analyse the variations of these proportions. For instance, darker regions will have high slopes of information variations whereas lighter regions will have slower variations (e.g., specular highlights).

For the opponent colour space, assuming dichromatic reflection and white illumination, the channels O_1 and O_2 are independent to highlights.

$$\begin{aligned}\begin{pmatrix} O_1 \\ O_2 \end{pmatrix} &= \begin{pmatrix} \frac{(m_b(\rho)C_b^R+m_i(\rho))-(m_b(\rho)C_b^G+m_i(\rho))}{\sqrt{2}} \\ \frac{(m_b(\rho)C_b^R+m_i(\rho))+(m_b(\rho)C_b^G+m_i(\rho))-2(m_b(\rho)C_b^B+m_i(\rho))}{\sqrt{6}} \end{pmatrix} \\ &= \begin{pmatrix} \frac{m_b(\rho)C_b^R-m_b(\rho)C_b^G}{\sqrt{2}} \\ \frac{m_b(\rho)C_b^R-m_b(\rho)C_b^G-2m_b(\rho)C_b^B}{\sqrt{6}} \end{pmatrix}\end{aligned} \tag{4.14}$$

O_1 and O_2 are still dependent on $m_b(\rho)$ and so they are sensitive to geometry, shading and the intensity of the light source. The O_3 channel is the intensity and contains no invariance properties at all.



O_1 O_2

Figure 4.7: The opponent colour channels.

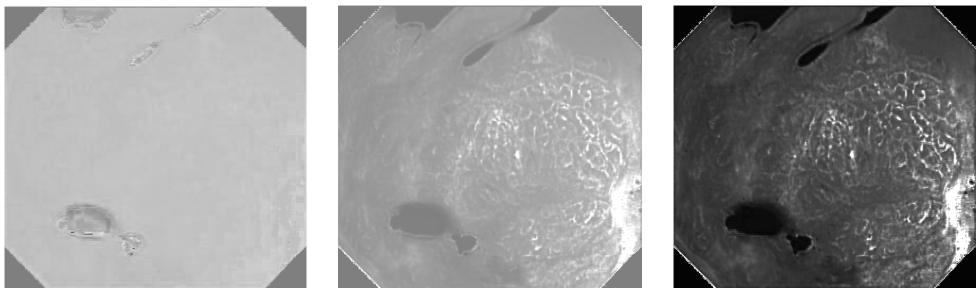
By taking the hue for the opponent colour space and assuming a matte surface we obtain:

$$hue = \arctan\left(\frac{O_1}{O_3}\right) = \arctan\left(\frac{\sqrt{3}(C_b^R - C_b^G)}{C_b^R + C_b^G + C_b^B}\right) \tag{4.15}$$

The chromatic opponent colours can also be computed.

$$\begin{pmatrix} C_a \\ C_b \end{pmatrix} = \begin{pmatrix} \frac{O_1}{O_3} \\ \frac{O_2}{O_3} \end{pmatrix} = \begin{pmatrix} \sqrt{\frac{3}{2}}\left(\frac{C_b^R-C_b^G}{C_b^R+C_b^G+C_b^B}\right) \\ \sqrt{\frac{1}{2}}\left(\frac{C_b^R+C_b^G-2C_b^B}{C_b^R+C_b^G+C_b^B}\right) \end{pmatrix} \tag{4.16}$$

Both the hue and the chromatic opponent colours are invariant to lighting geometry and specularities (Gevers et al., 2012).



hue C_a C_b

Figure 4.8: The Hue and the chromatic opponent colours.

Chapter 5

Pathology Recognition

We saw in Chapter 3 how to extract features from the images and based on the DRM we obtained special photometric invariant features in Chapter 4. In Section 2.2 the images were grouped into two classes according to Singh's taxonomy. But know a question arises. How can we actually build a CAD system to help us automatically classify these images? For this matter we will resort to a powerful classification tool known as Support Vector Machine, a learning mechanism that became very popular two decades ago. For self-contained reasons of this document we will review the concepts behind SVMs.

5.1 Linearly Separable Binary Classification

The design of automatic learning algorithms is one classic research problem from the pattern recognition community. The most well-known problem is classification. The idea is to take some input vector x and assign it to some discrete class, C_k where $k = 1, ...K$. The input feature space will be divided into decision regions, bounded by a called decision boundary. If the data set has classes that can be separated exactly by a computed linear decision boundary, the classes are said to be linearly separable. The simplest way of doing this is to construct a linear discriminant function that takes an input vector x and directly assigns it to a specific class. It is represented by the following expression:

$$g(x) = w^T x + w_0 \tag{5.1}$$

where w is the weight vector and w_0 is a bias parameter (Bishop & Nasrabadi, 2006). Reducing the problem to two classes, say C_1 and C_2 and assuming we have an input vector x, the linear discriminant function will assign it to class C_1 if $g(x) > 0$ and to class C_2 otherwise. The decision boundary is defined as $g(x) = 0$. If the feature space is D-dimensional, the decision boundary will be (D+1)-dimensional.

For any two points, x_a and x_b lying on the decision surface, we have that $w^T x_a + w_0 = w^T x_b + w_0$, in other words $w^T(x_a - x_b) = 0$, which indicates that w is normal to any vector lying in the decision boundary and thus defines the orientation of the decision surface.

Let us assume for simplicity that we have only two features, x_1 and x_2. Any point in the feature space will be a linear combination of these two features. The discriminant function gives us the decision boundary, $g = 0$, thus separating the feature space into two regions, R_1 and R_2. Considering that the distance of x to the decision boundary is r, as shown in Figure 5.1.

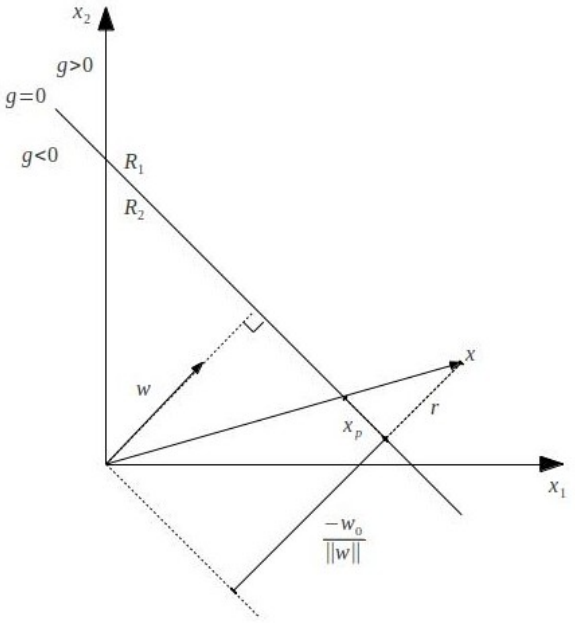


Figure 5.1: Linear discriminant function in a two dimensional feature space (Adopted from Bishop & Nasrabadi (2006)).

As x_p is the projection of x into the decision boundary, we can express x as:

$$x = x_p + r\frac{w}{\|w\|} \tag{5.2}$$

But x_p lies on the decision boundary and therefore $g(x_p) = 0$, so multiplying both sides of the equation by w^T and adding w_0, we obtain (Bishop & Nasrabadi, 2006):

$$g(x) = w^T x + w_0 = r\|w\| \tag{5.3}$$

which gives

$$r = \frac{g(x)}{\|w\|} \tag{5.4}$$

The linear discriminant thus divides the feature space through a decision boundary or hyperplane if we have more than two dimensions. The orientation of the hyperplane is controlled by the weight vector w and the bias parameter w_0 controls the location of the hyperplane relative to the origin of the feature space. The linear discriminant function $g(x)$ gives us a signed measure of the perpendicular distance of x to the decision boundary as demonstrated by Equation 5.4 (Bishop & Nasrabadi, 2006).

SVM rely on the same principles, but they represent the data in a much higher dimension than the original feature space by means of a linear mapping, $\phi(x)$.

$$g(x) = w^T \phi(x) + b \tag{5.5}$$

where b is the bias parameter. Lets assume, as before, we have N input vectors $x_1, x_n, ..., x_N$ with the corresponding labels $y_1, y_n, ..., y_N$ where $y_n \in \{-1, 1\}$. The transformed data points by means of $\phi(x)$ will be classified according to the sign of $g(x)$. Lets also assume the case where the training data set is linearly separable, this is, Equation 5.5 has at least one solution, satisfying the condition $g(x_n) \geq 0$ for points with label $y_n = +1$ and $g(x_n) < 0$ for points with $y_n = -1$. But multiple solutions may arise. We must then find the solution with the smallest generalization error (Bishop & Nasrabadi, 2006).

SVM makes use of the called *Support Vectors*, the points that are closest to the decision boundary. This distance is called the *margin* and the idea behind the SVM is to maximize this margin.

In Figure 5.2 we have two classes represented by the green and blue dots and the respective support vectors represented by a yellow contour. $x_1^{'}$ and $x_2^{'}$ are the transformed features by means of $\phi(x)$. m_1 and m_2 are the respective distances to the hyperplane separating the two classes. To implement a SVM we need to calculate the variables w and b so that our training

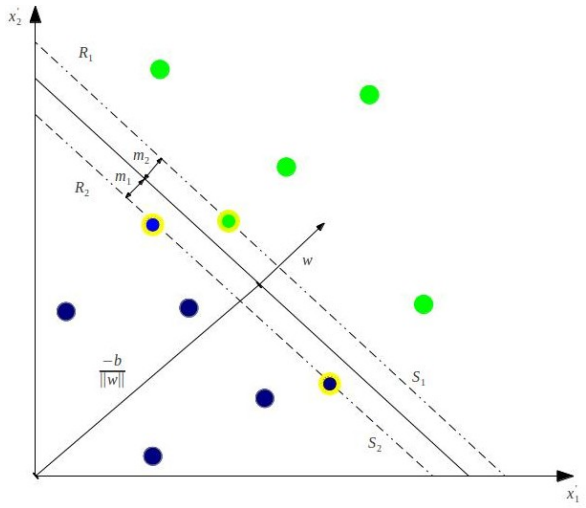


Figure 5.2: Decision plane through two linearly separable classes.

data can be described by:

$$\begin{aligned} w^T\phi(x_n) + b \geq 1 \quad \text{for} \quad y_n = +1 \\ w^T\phi(x_n) + b \leq -1 \quad \text{for} \quad y_n = -1 \end{aligned} \tag{5.6}$$

These equations can be combined into:

$$y_n(w^T\phi(x_n) + b) - 1 \leq 0 \quad \forall n \tag{5.7}$$

Considering just the support vectors, we can define two corresponding hyperplanes, S_1 and S_2 as:

$$\begin{aligned} w^T\phi(x_n) + b = +1 \quad \text{for} \quad S_1 \\ w^T\phi(x_n) + b = -1 \quad \text{for} \quad S_2 \end{aligned} \tag{5.8}$$

A SVM will find a hyperplane that makes the distances from the support vectors to the corresponding hyperplanes equal, thus defining the margin. The orientation of the hyperplane will be such, that the margin is maximized. By vector geometry we can find that the margin is equal to $\frac{1}{\|w\|}$ and maximizing it subject to the constraints defined by Equation 5.7 is equivalent to finding:

$$\min\|w\| \quad \text{such that} \quad y_n(w^T\phi(x_n) + b) - 1 \leq 0 \quad \forall n \tag{5.9}$$

Minimizing $\|w\|$ is equivalent to minimizing $\frac{1}{2}\|w\|^2$ and we need to perform a Quadratic Programming optimization. For more details see Bishop & Nasrabadi (2006).

5.2 Binary Classification for Data not Fully Linearly Separable

If the data is not fully linearly separable, we need to relax the constraints slightly in order to allow for misclassified points. This is achieved introducing a positive slack variable, ξ_n as illustrated in Figure 5.3.

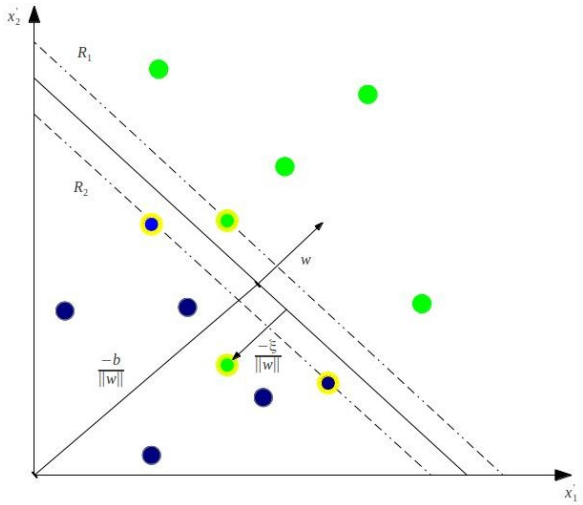


Figure 5.3: Decision plane through two non-linearly separable classes

In this case the equations for the hyperplanes will have to account for the penalty factor and so they can be written as:

$$\begin{aligned} w^T\phi(x_n) + b &\geq +1 + \xi_n \quad \text{for} \quad y_n = +1 \\ w^T\phi(x_n) + b &\leq -1 + \xi_n \quad \text{for} \quad y_n = -1 \\ \xi &\geq 0 \quad \forall n \end{aligned} \tag{5.10}$$

which can be combined into:

$$y_n(w^T\phi(x_n) + b) - 1 + \xi_n \geq 0 \quad \text{where} \quad \xi_n \geq 0 \forall n \tag{5.11}$$

This is called a soft margin SVM. Decision points on the incorrect side of the decision boundary will have a penalty factor that increases with the distance from it. To try to reduce the number of miss-classifications is to find:

$$\min \frac{1}{2}\|w\|^2 + C\sum_{n=1}^{N}\xi_n \tag{5.12}$$

subject to the constraints:

$$y_n(w^T\phi(x_n) + b) - 1 + \xi_n \geq 0 \quad \forall n \tag{5.13}$$

where the parameter C controls the trade-off between the slack variable penalty and the size of the margin. See Bishop & Nasrabadi (2006) for more details.

5.3 Non-linear Support Vector Machine

When applying our SVM to linearly separable data we create a matrix H, from the dot product of our input variables.

$$H_{n,m} = y_n y_m K(x_n, x_m) = x_n \cdot x_m = x_n^T x_m \tag{5.14}$$

$K(x_n, x_m)$ is an example of a family of functions called *Kernel functions*, where $K(x_n, x_m)$ is the *Linear kernel*. There are other kernel functions as the *Intersection kernel* defined as $K_{HI}(x_n, x_m) = \min(x_n, x_m)$ or the χ^2 *kernel*, $K_{\chi^2} = \frac{2x_n^t x_m}{x_n + x_m}$. These special functions allow to transform the features into a higher dimensional space, so that the data can actually be completely separated and we can classify our images based on the constructed decision plane. We shall not enter in many details as this is a wide subject and is out of the scope of this thesis[1].

[1] For a more detailed description on this subject see Shawe-Taylor & Cristianini (2004)

Chapter 6

Experimental Study

We begin this chapter by presenting the image dataset used in these experiments and the separation that we performed in order to build a training set and a test set. We also present the used methodology to extract features from the images in order to properly recognize lesions in the tissues. The method consists of extracting SIFT features, sampling the descriptors with a regular grid. The method to construct new image histograms is also presented forming the *vocabulary* of each image. These new histograms will be the input for our classifier. As we saw in Chapter 5 the classification step will be performed using a Support Vector Machine (SVM). The methods that we used in order to optimize the classification process and to assess the classifier performance are also presented.

6.1 Image dataset

The image set is composed of endoscopic images from the Barrett's oesophagus, collected by a research group. We have a total of 250 images classified according to Singh's classification proposal (Singh et al., 2008) as presented in Section 2.2. They were also annotated, this is, the region of interest was selected in order to discard irrelevant information, by clinical experts with different expertises. In this thesis the images were grouped into two classes, normal (C_1) versus pathologic (C_2), reducing the problem to a binary classification system. Type A images are normal, they do not present any kind of irregularity either in the pit patterns or in the microvasculature. Types B, C and D do show some irregularities and may

be evolving to cancer. Having this in mind Type A is from now on relative to C_1 and the other types to C_2, as illustrated in Table 6.1. From the 250 images we have 61 images of class C_1 and 189 images of class C_2. As we can see the dataset is quite unbalanced.

Singh's Classification	Number of Images	Assigned Class
Type A	61	C_1
Type B	81	C_2
Type C	17	
Type D	61	

Table 6.1: Grouping the images for a binary classification system.

The separation in two classes is thought to separate normal images from images that may evolve to cancer and therefore require some deeper analysis. In order to train our SVM we separated the images into a training dataset and a test dataset. The first one is used to train the classifier and the second one to evaluate the classifier performance. To balance the dataset in the training stage we chose to use 50 images of class C_1 and 50 images for class C_2 leaving 11 of class C_1 and 139 of class C_2 for the test dataset.

6.2 Methodology

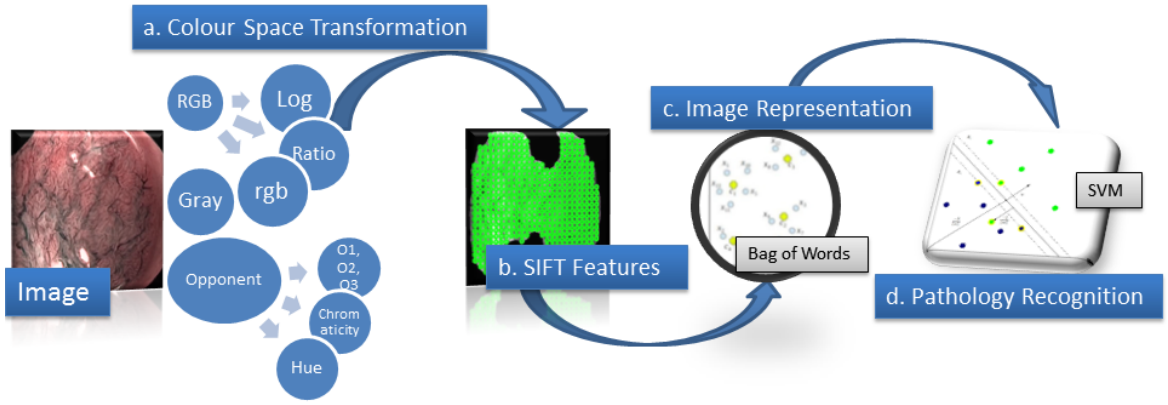


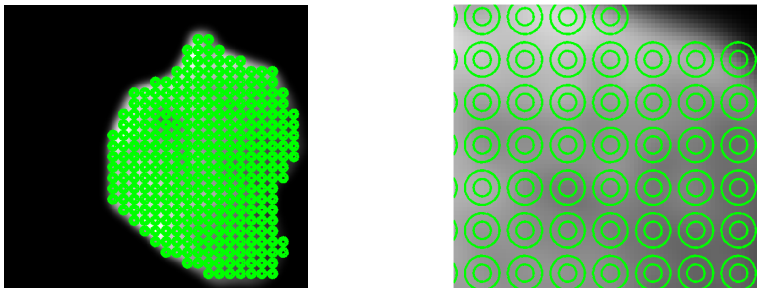
Figure 6.1: Experimental Setup

In this Section we will describe the method to build our framework. Figure 6.1 illustrates the methodology for this purpose.

6.2.1 Extracting colour features

In this thesis SIFT features were extracted using a dense approach since the texture of NBI images fills the whole image. The images were previously annotated and thus we rejected the key-points or frames lying outside the region of interest. Figure 6.1 shows the methodology used in this thesis. The extraction of SIFT features (step b), the representation of the images (step c) and the recognition of the pathology (step d), were performed using the vlfeat library (see http://www.vlfeat.org/ for a detailed description).

Figure 6.2 illustrates an example of the application of the DSIFT approach to a masked image. Two different scales were also used to extract the descriptors as seen by the two different radius circles.



(a) Application of DSIFT to a masked image. (b) The two scales.

Figure 6.2: Multi-scale DSIFT extraction.

From Figure 6.3 we see that there are several parameters that can be adjusted like the sampling step of the key-points and the bin size of the descriptor. We fixed these values in 10 for the sampling step and 8 for the bin size as these gave the best empirical results.

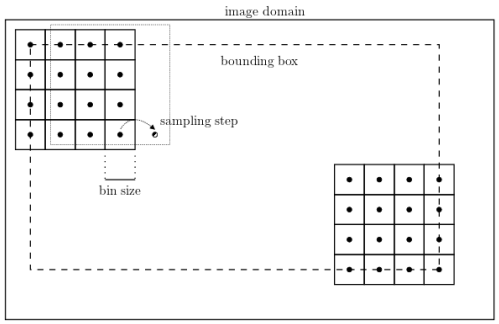


Figure 6.3: The geometry of the DSIFT approach.

The sampling step controls the horizontal and vertical distance between the key-points and the bin size controls the number of pixels covered by the spatial bin of the descriptor. In fact the size of the spatial bin is related to the key-point scale by a magnification factor, so the descriptors are extracted at a scale equal to $S = \frac{\text{bin size}}{\text{magnification factor}}$. The image is also pre-smoothed with a Gaussian of variance $(S)^2 - 0.25$ (Vedaldi & Fulkerson, 2008).

The methodology to study the impact of colour information begins with the SIFT features extracted converting the image to gray. To add information from colour and evaluate the possible gains we incrementally add colour information following the photometric invariant properties presented in Chapter 4.For this purpose we extracted features from the R, G and B channels independently and added them to the gray descriptor, obtaining the set $\{gray; rgb\}$. Adding the set of first-order colour invariants $\left\{\frac{R}{G}; \frac{R}{B}; \frac{G}{B}\right\}$ to the previous set we get $\{gray; rgb; ratio\}$. We then took the logarithm of each ratio and added to all the other descriptors. For simplicity this means $\{gray; rgb; ratio; log\}$. The descriptors from the opponent colour space were also added. As for the RGB, the descriptors are extracted independently for each channel $\{O_1; O_2; O_3\}$ adding them to the previous ones, $\{gray; rgb; ratio; log; opp\}$ In a final step we extracted the features from chromaticity opponent channels given by Equations 4.16. Simultaneously we also extracted features from the hue and added them to the previous descriptors, obtaining the set of descriptors $\{gray; rgb; ratio; log; opp; chroma; hue\}$. All the descriptors were stacked vertically in order to increase dimensionality.

Algorithm 6.1 Method to extract multi-scale features from the images

Define $samplingstep = 10; binsize = 8; magnification = \{2, 4\}$;

Convert the image to gray

for n=1 to N **do**

  for i=1 **do** 2 ▷ We used 2 scales for smoothing

    Calculate the $scale_i = \frac{binsize}{magnification_i}$

    Smooth the image with a Gaussian of variance: $(scale_i)^2 - 0.25$

    Extract the frames and the descriptors for $image_n$

  if *frames outside the mask* **then**

    Discard frames and descriptors

  repeat the algorithm with the next set of descriptors

6.2.2 Building the vocabulary

Bag of Words: Each extracted key-point is then a vector in the feature space described by a histogram of oriented gradients. An image will be described by a set of these descriptors that can be thought as a set of *visual words* where each key-point is described by a *visual word*. If we group these *visual words* using a clustering algorithm such as k-means, we are grouping the determined *visual words* and describing the image with more meaningful words. These new words are called the *visual terms*. By counting the frequency of the *visual words* where the *visual terms* serve as bins, we can build a histogram that describes the image. This histogram is called *visual vocabulary*.

K-means: K-means consists of a simple unsupervised approach, meaning that the labels or the classes of each image are never present in this stage. As a brief description lets suppose we have a set of data points $\{x_n\} \in \mathbb{R}^d$, collected from all the images and where d is the dimension of the feature space, in this case $d = 128$. K-means will search for K cluster centres, the vectors $\{c_k\} \in \mathbb{R}^d$ and will also assign the data points to this centres, through a function $A : \{x_n\} \rightarrow \{c_k\}$. The goal is to find an assignment of data points to clusters and a set of vectors $\{c_k\}$ such that the sum of the squares of the distances to each data point to its closest vector c_k is minimum (Vedaldi & Fulkerson, 2008).

The Vocabulary: For each image individually the extracted data points will be assigned to the calculated centres, or *visual terms*. By counting how many points of each image we have near the *visual term* we build a histogram which is our *visual vocabulary* for each image of the form $H_n = [h_1^n, h_2^n, h_3^n, ..., h_k^n]$, as illustrated in Algorithm 6.2, where n is the image number, K is the number of clusters and h are the counts for each *visual term*.

6.2.3 Training the classifier

The training of the classifier is done using an SVM, see Chapter 5, using only the training dataset. In order to assess the overall performance, we now used the test dataset, trying to place these never *seen* images into one side of the decision hyperplane. It is important to refer that the input vectors for our SVM will be the *vocabulary* calculated by the k-means

Algorithm 6.2 Calculating the *visual terms* with k-means.

Initialize $\{c_k\}$;

for i=1 to K **do**

Assign each point to nearest c_k according to:$\|x_n - c_i\|^2$

recompute c_i

repeat the assignments to nearest center

until $\sum \|x_n - c_i\|^2$ is minimum

return $\{c_1, c_2, ..., c_k\}$

Algorithm 6.3 Building the *vocabulary* for each image

for n=1 to N **do**

find $A_n : \{x_n\} \rightarrow \{c_k\}$

count number of points associated to each cluster

compute $H_n = \{h_k^n\}$

method. The dimensionality of the space is now equal to the number of *visual terms* we use. A sub-gradient solver for a SVM was used to compute the penalty factor, defined in Equation 5.12, the so called *hinge loss* function used for maximum-margin classifiers (Vedaldi & Fulkerson, 2008).

6.3 Validating the parameters

In order to evaluate the best regularization parameter, $\lambda = \frac{1}{C}$, we performed a three-fold cross validation on the training data. We created three subsets of the training data by taking approximately $\frac{1}{3}$ of the images from each of the two classes. Two of them are grouped and they are responsible for the training stage and the other is used as a validation set. The best λ is chosen by considering the combination that offers the minimum error which is calculated by:

$$error^{fold} = \frac{1}{N^{val}} \sum_{i=1}^{N^{val}} I(y_i^*; y_i) \quad with \quad I(y_i^*; y_i) = \begin{cases} 1 & \text{if} \quad y_i^* \neq y_i \\ 0 & \text{otherwise} \end{cases} \tag{6.1}$$

where N^{val} is the number of validation images, y_i are the labels with $*$ denoting the predicted labels. Ten possible values for λ were tried ranging from 0.1 to 1. For each λ the error is

calculated and once the best λ is determined a mean error is also calculated considering the mean value of the errors obtained in the three possible combinations for the corresponding value of λ.

Algorithm 6.4 3-fold Cross-Validation

for λ= 0.1:0.1:1 **do**

for *fold=1:3* **do**

$foldtrain = [1, 2, 3] \backslash fold$

Build Vocabulary *for the training and test folds*

Create *Labels for the training and test folds according to:*

$\{C_1\} \rightarrow -1 \quad \{C_2, C_3, C_4\} \rightarrow 1$

Build *Model by training the classifier*

Test *The classifier with the validation set*

Determine *Minimum error given by Equation 6.1 and the corresponding* λ

For that λ*:*

return *Mean value for the three combinations*

6.4 Assessing the Classifier Performance

For the assessment of the classifier performance we used the Mean Average Precision (MAP) as a parameter. The MAP is calculated taking the mean value of the precision which in turn can be calculated by the error given in Equation 6.1. The advantage of using the MAP lies on the fact that we have a very unbalanced test set and the MAP allows to penalize more representative classes. It is given by:

$$MAP = \frac{1}{K}\left[\sum_{j=1}^{K}\frac{1}{N_j}I(y_i^*; y_{i,K})\right] \tag{6.2}$$

where K in the number of classes, N is the number of images and y_i are the labels with $*$ denoting the predicted labels. The fact that the MAP performs this division by the number of classes, in this case two, will penalize the most representative class, in this case class C_2.

The methodology presented in Algorithm 6.5 was repeated 10 times in order to obtain more stable MAP values. To save computational time the extraction of the features can be

done only at the first run. The results of this experiments and others as to determine the best regularization parameter, the number of *words* to use for our *vocabulary* and the more adequate kernel to use in our classifier, are presented and discussed in detail in Chapter 7.

Algorithm 6.5 The Classifier

Extract *Features* $\forall n$ *according to Algorithm 6.1*

Build Vocabulary $\forall n$ *according to Algorithm 6.2*

Create *Training set according to Section 6.1*

Label *according to*

$\{C_1\} \rightarrow -1 \quad \{C_2, C_3, C_4\} \rightarrow 1$

Determine *Regularization parameter* λ *performing a 3-fold cross-validation according to Algorithm 6.4*

Build *Model with the training set*

Test *Classifier performance with the test set*

Return *MAP*

Chapter 7

Results and Discussion

There are several parameters that need to be determined before assessing the impact of the colour descriptors on the performance of the classifier. One of them is the type of kernel to use in our SVM. For this purpose we performed a few quick tests considering the Linear, Intersection and the χ^2 kernel, presented in Section 5.3. The tests consisted in analysing the mean error obtained in the cross-validation procedure given by Equation 6.1, for vocabulary sizes of 200, 500 and 1000. The results are presented in Table 7.1. The tests are merely qualitative and so they do not intend to be a justification on the type of kernel to use for NBI images. For a more detailed study on this subject see Sousa et al. (2013) and Maji et al. (2008).

	Linear Kernel		Intersection Kernel		χ^2 Kernel	
# clusters	error	λ	error	λ	error	λ
200	0.2143	0.2	0.2123	0.2	0.1667	0.2
500	0.2326	0.4	0.2093	0.1	0.3256	0.6
1000	0.2193	0.3	0.0932	0.7	0.2326	0.4
$\bar{e}$	0.2237		0.1716		0.2416	

Table 7.1: Mean errors obtained for the Linear, Intersection and χ^2 kernels.

Considering the mean value of the error, ($\bar{e}$), for the three vocabulary sizes we can see that the Intersection kernel was the one that gave the least error. For this reason we used in our experiments an Intersection kernel for the SVM.

Another parameter of extreme importance is the size of the visual vocabulary. With the intersection kernel, we performed some tests trying to assess the best number of visual words to use. It seems reasonable that increasing the vocabulary size will increase the classifier performance but that will also increase the computational time of the experiments. So a compromise between performance and computational cost has to be achieved.

# of visual words	MAP	Δt(s)
100	0.7328	493
200	0.7716	906
400	0.7944	1437
800	0.7842	1636
1600	0.7980	2342

Table 7.2: MAP vs # of visual words vs Δt.

The qualitative tests in Table 7.2, that result from a medium value of the MAP and the $\Delta t(s)$ for three runs, indicate that the MAP seems to converge to an optimal value after 400 clusters. For bigger values the computational cost is not justified. We therefore used for our tests 400 visual words in order to build our visual vocabulary.

To perform our tests, each run will consist of extracting the SIFT features from the images, build the vocabulary with 400 visual words and evaluate the classifier performance with an intersection kernel. In order to obtain a stable MAP result we repeated our experiment 10 times for each set of descriptors, that are described in Section 6.2. Table 7.3 shows the obtained results.

Descriptors \ Run #	1	2	3	4	5	6	7	8	9	10
{*gray*}	0.726	0.814	0.863	0.777	0.819	0.796	0.756	0.808	0.774	0.811
{*gray*; *rgb*}	0.883	0.753	0.899	0.770	0.757	0.881	0.845	0.901	0.721	0.783
{*gray*; *rgb*; *ratio*}	0.729	0.776	0.818	0.799	0.872	0.863	0.759	0.857	0.808	0.871
{*gray*; *rgb*; *ratio*; *log*}	0.820	0.838	0.857	0.808	0.780	0.859	0.836	0.753	0.708	0.767
{*gray*; *rgb*; *ratio*; *log*; *opp*}	0.794	0.879	0.823	0.865	0.771	0.857	0.843	0.808	0.843	0.914
{*gray*; *rgb*; *ratio*; *log*; *opp*; *chroma*; *hue*}	0.797	0.788	0.848	0.707	0.775	0.842	0.823	0.815	0.819	0.892
{*opp*}	0.547	0.787	0.727	0.752	0.811	0.794	0.836	0.798	0.685	0.744

Table 7.3: The results of the ten simulations for each descriptor set.

It can be seen that the combination of $\{gray; rgb; ratio; log; opp\}$ attains almost always the best results. The set $\{gray; rgb\}$ seems to have also good results, in some of the runs was actually the best, the fluctuations around the mean value are greater then the set $\{gray; rgb; ratio; log; opp\}$, as it can be observed by Table 7.4 analysing the standard deviations between both sets. The high fluctuations of these results in each run is mostly due to stochastic behaviour of the K-Means algorithm.

Descriptors	MAP
$\{gray\}$	0.794 ± 0.038
$\{gray; rgb\}$	0.819 ± 0.069
$\{gray; rgb; ratio\}$	0.815 ± 0.050
$\{gray; rgb; ratio; log\}$	0.805 ± 0.049
$\{gray; rgb; ratio; log; opp\}$	0.840 ± 0.042
$\{gray; rgb; ratio; log; opp; chroma; hue\}$	0.811 ± 0.049
$\{opp\}$	0.748 ± 0.083

Table 7.4: The impact of colour on the classifier performance.

We can also observe from Table 7.4 that the addition of colour information and photometric invariant combinations of the colour channels also improves the results. This improvement could be explained by the adding of discriminative power to our classifier, this is, we are adding more dimensions, so it seems reasonable to say that the classifier performance would naturally increase by this increase in dimensionality. But as we can observe the addition of the logarithmic transformation of the ratio of the channels, the set $\{gray; rgb; ratio; log\}$, performs actually worse then with the set $\{gray; rgb; ratio\}$. Although the ratio of the channels offers photometric invariances to the surface orientation and illumination direction and intensity, the adding of non-linear transformations, as the logarithm tends to intensify the amount of noise so a small perturbation of the RGB values causes a large jump in the transformed values. When we add the *chroma* and the *hue* to the previous set with the opponent colours the performance also decreases. Again, recall that both the *chroma* and the *hue* are computed by ratios of the opponent colours and this may introduce noise. So the increase in performance is not due to the addition of more dimensions by taking more SIFT features but is mostly due to the addition of useful photometric invariant colour information.

From Table 7.4 we also observe that the set $\{gray; rgb; ratio; log; opp\}$ gives the highest MAP. This improvement in the results is mainly due to the fact that the opponent channels highly decorrelate the RGB channels. For example, the O_1 channel performs the difference of the R and G channels, Equation 4.6. By doing so we are reducing possible redundant information in the R and G channel. We have also studied the performance of the classifier by considering just the descriptors extracted independently from each opponent channel, the set $\{opp\}$. The results were actually worse then with the set $\{gray\}$. This fact could be due to the lack of photometric invariances in the set of descriptors of just the opponent colours. Recall that the O_3 channel is simply the intensity and it reveals no invariance properties at all as for the gray-scale image that possesses the constancy of the RGB colours. We can also observe from Table 7.4 that the results for the set $\{opp\}$ is the one that reveals the highest standard deviation value for all experiments. Once again the lack of some photometric invariant properties could explain these high standard deviations allied to the k-means process.

In order to correctly visualize the performance of the classifier with each set of descriptors it is usually used in machine learning a confusion matrix. Each column represents the instances in a predicted class and each row represents the instances in an actual class thus representing a visual analysis if the classifier is confusing the two classes. For our binary problem we assigned class C_1 as a negative and class C_2 as positive.

		Predicted	
		C_2	C_1
Real	C_2	TP	FN
	C_1	FP	TN

Table 7.5: A typical confusion matrix for a binary problem.

A *true positive* (TP) is an image predicted as C_2 and is in fact C_2. A *false positive* (FP) is an image predicted as C_2 but is actually C_1. A *false negative* (FN) is an image assigned to class C_1 but is of class C_2 and finally a *true negative* is a class C_1 correctly classified. With the separation of our dataset as indicated in Section 6.1, we then have 150 images to test our classifier. The tests were performed following Algorithm 6.5 presented in Section 6.4. The results in Table 7.6 show the confusion matrices for each set of descriptors tested, considering the mean values of TP, FP, FN and TN for the ten runs.

105	34
2	9

(a) $\{gray\}$

109	30
2	9

(b) $\{...; rgb\}$

107	32
2	10

(c) $\{...; ratio\}$

109	30
2	9

(d) $\{...; log\}$

110	29
1	10

(e) $\{...; opp\}$

109	30
2	9

(f) $\{...; chroma; hue\}$

102	37
3	8

(g) $\{opp\}$

Table 7.6: Confusion matrices for each set of descriptors in average for the 10 simulations.

In order to analyse and interpret the information from these confusion matrices there are a few parameters that are usually used. These parameters allow a better evaluation of our classifier performance.

First we have have the *accuracy* (AC) of the classifier which is defined as the proportion of the total number of correct predictions $(TP + TN)$:

$$AC = \frac{TP + TN}{TP + TN + FP + FN} \tag{7.1}$$

The recall or *true positive rate* (TPR) is the proportion of positive images correctly classified:

$$TPR(sensitivity) = \frac{TP}{TP + FN} \tag{7.2}$$

This rate is also know as the *sensitivity* of the classifier.

Next we have the *false positive rate* (FPR) defined as the proportion of the negatives incorrectly classified:

$$FPR = \frac{FP}{FP + TN} \tag{7.3}$$

The *true negative rate* (TNR), or *specificity*, is defined as the proportion of negative classes correctly classified:

$$TNR(specificity) = \frac{TN}{FP + TN} \tag{7.4}$$

The *false negative rate* (FNR) is the proportion of positive classes incorrectly classified:

$$FNR = \frac{FN}{TP + FN} \tag{7.5}$$

And finally the *precision* (PR), defined as the proportion of correctly predicted classes:

$$PR = \frac{TP}{TP + FP} \tag{7.6}$$

In Table 7.7 we present all of these rates calculated for each set of descriptors.

Parameters / Descriptors	AC	TPR	FPR	TNR	FNR	PR	MAP
$\{gray\}$	0.759	0.753	0.164	0.836	0.247	0.983	0.794±0.038
$\{gray; rgb\}$	0.789	0.784	0.145	0.855	0.216	0.986	0.819±0.069
$\{gray; rgb; ratio\}$	0.774	0.767	0.136	0.864	0.233	0.986	0.815±0.050
$\{gray; rgb; ratio; log\}$	0.785	0.782	0.173	0.827	0.218	0.983	0.805±0.049
$\{gray; rgb; ratio; log; opp\}$	0.796	0.788	0.109	0.891	0.212	0.989	0.840±0.042
$\{gray; rgb; ratio; log; opp; chroma; hue\}$	0.789	0.785	0.164	0.836	0.215	0.984	0.811±0.049
$\{opp\}$	0.735	0.732	0.236	0.764	0.268	0.975	0.748 ± 0.083

Table 7.7: Evaluation of several parameters for each set of descriptors.

We observe that the set of descriptors that attained the best results for all the parameters was the set $\{gray; rgb; ratio; log; opp\}$. It attained the highest *accuracy* which means it had the highest number of correct predictions. It also attained the highest *sensitivity* value so the proportion of positive images correctly classified was also the highest. It also attained the lowest FPR and FNR. In this case lower is better because these rates actually measure the proportion of miss-classifications. The highest *specificity* of this set of descriptors tells us that it was the best in classifying the negative images. Finally the highest values for the *precision* and for the MAP were also for this set of descriptors thus attaining the highest proportion of correctly predicted classes. The *precision* was actually high for all the sets of descriptors but is due to the fact that our test set is quite unbalanced. We have more positive then negative images. This is a parameter that is actually biased to the data and for this matter the MAP is a more trusty measure because it weights the set according to the number of images per class.

Probabilistically speaking the *sensitivity* or the TPR can be viewed as the probability of a certain image to be of class C_2, this is positive, knowing that it is of class C_2. In other words it means that if we have an image that was classified as a positive we have *sensitivity*% shore that it was well classified. If a classifier is highly sensitive, this is the TPR is close to 1, the number of images of class C_2 that go undetected decreases. For the *specificity*, or the TNR, it can be viewed as the probability that an image is of class C_1 given it is of class C_1. If an image is classified as negative, this is, no cancer, we have *specificity*% shore it is a correct

classification. The higher the *specificity*, fewer images of class C_1 will be miss-classified as class C_2.

For a binary classifier as is ours the *sensitivity* and *specificity* are usually sufficient parameters to assess the classifier performance, plus the MAP. So for the set that shows the best results, the set $\{gray; rgb; ratio; log; opp\}$, we present these parameters in Table 7.8 in order to enhance the classifier performance build with this set.

Descriptor set	*sensitivity*	*specificity*	MAP
$\{gray; rgb; ratio; log; opp\}$	0.788	0.891	0.840±0.042

Table 7.8: Binary classifier performance assessment

The use of the set $\{gray; rgb; ratio; log; opp\}$ for the SIFT feature extraction process, the use of the *Bag of Words* method to cluster the features and the use of the SVM with an Intersection Kernel thus constitutes our proposal for a robust framework to perform classification on NBI images of the oesophagus.

Chapter 8

Conclusions and Future Works

In this thesis we proposed to develop a robust framework for the classification of NBI images of the oesophagus. We used a local descriptor densely sampled, DSIFT, to describe these new images, performed the quantification of these descriptors using the *Bag of Words* method and used a SVM with an intersection kernel for the classification.

The main goal was to study the impact of colour information to the base SIFT descriptor. Adding this information was performed by following some physical properties of the images itself and modulating them with the Dichromatic Reflection Model that considers the tissue as being composed by two layers. This modulation allows to rebuild the information on the RGB channels and in turn obtain special photometric invariants. To the invariance of the SIFT descriptor to scale, rotation and translation, we were able to add information invariant to viewpoint, lighting geometry, specularities, surface orientation, illumination direction and intensity and shading.

The obtained results show that to add information from the RGB channels independently, the ratio of the RGB channels, the logarithmic of these ratios and the information from the opponent colour gives the best results. With this set of local descriptors we attained a MAP of 84.0%, a *sensitivity* of 78.8% and a *specificity* of 89.1%. Its invariance properties allied to the very special conditions that these images are acquired allow these descriptors to perform very well. The addition of colour information improved the classifier performance in about 5% comparing of course the referred set of local descriptors with the opponent colours to our

base gray descriptor. The fact that the opponent colour system highly decorrelates the RGB channels explains this gain in the performance.

The work presented in this thesis can actually be extended and improved. Some of the performed tests were merely qualitative as the tests performed to assess the more adequate kernel to use in our SVM. The intersection kernel was chosen among the χ^2 kernel and the linear kernel performing only three experiments comparing the mean error for three possible vocabulary sizes. With this is mind a more detailed study maybe performed in order to attain the best kernel method to use. Another qualitative test performed was the size of the vocabulary although the results do seem to justify the use of 400 visual terms to describe these images for the sake of computational cost. Again maybe a more detailed study on this subject may allow to determine a more effective size of the vocabulary. The use of these qualitative tests is justified by the main goal of this thesis that was to assess the classifier performance by adding colour information and so some parameters had to be fixed in order to perform the tests.

Another topic worthy of interest is the process of the image formation in the NBI system itself. A deeper knowledge of some of the characteristics of the NBI filters that are used and the process of rebuilding the image in the RGB monitor could provide valuable information on the degree of correlation of the channels and possibly one could modulate the images taking this information in consideration and obtain better results.

Appendix A

Abbreviations

CCD Charged-Coupled Device

NBI Narrow-band Imaging

RGB Red Green Blue

RORENO *Registo Oncológico Regional do Norte*

IPO-Porto *Instituto Português de Oncologia do Porto Francisco Gentil*

GERD Gastroesophageal Reflux Disease

SCC Squamous Cell Carcinoma

BE Barrett's Oesophagus

SIFT Scale Invariant Feature Transform

DSIFT Dense SIFT

DoG Difference-of-Gaussian

SVM Support Vector Machine

MAP Mean Average Precision

CAD Computer Aided Diagnosis

DRM Dichromatic Reflection Model

Appendix B

Published Article at RecPad2013

Colour Invariant Features for Narrow-Band Imaging in Gastroenterological Examinations

Bruno Mendes[2]
brunomendes81@gmail.com
Ricardo Sousa[1]
rsousa@dcc.fc.up.pt
Carla C. Rosa [2]
ccrosa@fc.up.pt
Miguel T. Coimbra[1]
mcoimbra@fc.up.pt

[1] IT, Department of Computer Science
FCUP
Porto, Portugal

[2] INESC TEC
UOSE - FCUP
Porto, Portugal

Abstract

In this work we developed a Computer-Aided Decision (CAD) system by making use of some physical properties in the image acquisition process and rebuilding the R, G and B channels achieving special photometric invariances. This is performed using a local descriptor such as the Scale Invariant Feature Transform (SIFT) and we assess the classifier performance by studying the impact of colour information to the descriptor. We achieved a performance of 79% with a regular gray conversion and 84% by making use of the opponent colours. The proposed set of descriptors achieved a *sensitivity* of 79% and a *specificity* of 89%.

Narrow-Band Imaging (NBI) is a promising tool in the diagnosis of cancer in gastroenterological images. It illuminates the mucosa with blue and green light and the fact that green penetrates deeper in the tissue will increase the contrast of superficial and deeper vessels and the visualization of certain structures that were unseen with conventional white light illumination. This brings new patterns that need to interpreted and analysed correctly to perform an accurate diagnostic. In this context a CAD Support System specialized for these images is crucial.

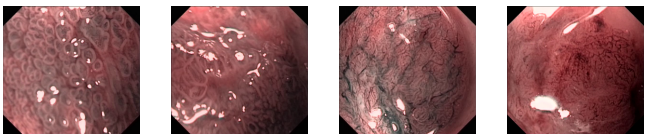
1 Introduction

NBI consists in narrowing the light output, illuminating the mucosa with blue and green with the use of special filters. Due to the different penetration depths of light (green penetrates deeper) and the fact that they match the absorption peaks of haemoglobin, blue will be absorbed by superficial vessels while green will be absorbed by deeper vessels thus enhancing the contrast between superficial and deeper vessels. Narrowing the bandwidth of the illumination also reduces scattering effects on the mucosal surface and thus the resulting image reveals structures and patterns that where unseen with a conventional white light illumination [3]. Blue light is the input for the B and G channels and so the superficial vessels and structures will appear brownish while green is the input for the R channel and so deeper vessels will appear with a cyan colour [3].

These new patterns have a high correlation with histology as shown by Singh et al. [6]. They proposed a grading system for these images based on the vascular and structural pattern observed at the oesophageal mucosal surface:

- Type A: Round pits with regular microvasculature;
- Type B: Villous/ridge pits with regular microvasculature;
- Type C: Absent pits with regular microvasculature;
- Type D: Distorted pits with irregular microvasculature;

Figure 1 shows example images from our data set. We reduced the problem to a binary classification system by assigning Type A to class C_1, the normal cases, and Types B, C and D to class C_2, the cases that show an evolution to cancer. The main idea is to build a CAD system capable of separating normal from abnormal cases with the use of local descriptors.



Type A Type B Type C Type D

Figure 1: Example images from our dataset.

2 Colour Invariant Features

2.1 Local Descriptor

In this work we propose to extract features using a local descriptor that shows invariance to scale, rotation and translation, the SIFT descriptor [2]. Patterns are spread over a wide region and for this reason the best way of sampling these images is to perform a dense sampling, as shown by Nowak et al. [4], placing a regular grid on top of the image and extracting the descriptors. The SIFT descriptor consists in computing the gradients in 8 possible orientations on an image patch of 4×4. The resulting vector is thus in a 128 dimensional feature space. The base descriptor is extracted with a regular conversion to gray-scale. In this work we intend to study the addition of colour information to this base descriptor and attend the possible improvements in the classification performance.

2.2 Adding Colour Information

The addition of colour to the base gray descriptor is performed by modulating the images according to the Dichromatic Reflection Model (DRM).

The Dichromatic Reflection Model The recorded image by a camera is the the sum of the incoming light at pixel position ρ, $L(\lambda,\rho)$, into the sensor, weighted by the spectral transmittance of the filter, $\tau(\lambda)$ and the spectral responsivity of the camera, $s(\lambda)$, over all wavelengths:

$$C = \int_{\lambda} L(\lambda,\rho)\tau(\lambda)s(\lambda)d\lambda \tag{1}$$

This process of spectral integration is a linear transformation [5]. The DRM considers, for an inhomogeneous material and neutral interface reflection, that the formed image of an object can be seen as a combination of two terms: *body* and *surface* reflection, for more information see [5], and we can then consider the incoming beam as colour triple $C = [R,G,B]$. The DRM states that:

$$C = m_i C_i + m_b C_b \tag{2}$$

where the indices i and b denote interface and body respectively, m_i and m_b are the magnitudes of the corresponding reflections, C_i and C_b are the corresponding colours and are also colour triples.

Photometric Invariants To obtain photometric invariants we rebuilt the R, G and B channels according to the DRM. For simplicity we assume a matte surface, where the term m_i responsible for the specular reflections is negligible. In doing so we obtain the Lambertian model for diffuse body reflection:

$$C = m_b C_b \tag{3}$$

A zero-order invariant can be obtained by normalizing the R, G, and B colours (*rgb*):

$$r = \frac{R}{R+G+B} = \frac{m_b C_b^R}{m_b(C_b^R + C_b^G + C_b^B)} = \frac{C_b^R}{C_b^R + C_b^G + C_b^B} \tag{4}$$

Similar equations can be obtained for the normalized G and B channels. By doing so we are obtaining invariance to lighting geometry and viewpoint [1]. A set of first-order invariants can be obtained by taking the ratio of the channels (*ratio*). They have the same invariant properties as the normalized RGB colours.

$$\left\{ \frac{R}{G}, \frac{R}{B}, \frac{G}{B}, \dots \right\} \tag{5}$$

In order to consider the quick density variations that cannot be captured linearly, we introduce the proportion under the logarithm (*log*) :

$$\left\{\log\left(\frac{R}{G}\right),\log\left(\frac{R}{B}\right),\log\left(\frac{G}{B}\right),\cdots\right\} \tag{6}$$

Human perception of colours relies on the opponent process theory. For this matter we consider the opponent colour space (*opp*):

$$\begin{pmatrix} O_1 \\ O_2 \\ O_3 \end{pmatrix} = \begin{pmatrix} \frac{R-G}{\sqrt{2}} \\ \frac{R+G-2B}{\sqrt{6}} \\ \frac{R+G+B}{\sqrt{3}} \end{pmatrix} = \begin{pmatrix} \frac{m_b(\rho)C_b^R - m_b(\rho)C_b^G}{\sqrt{2}} \\ \frac{m_b(\rho)C_b^R + m_b(\rho)C_b^G - 2m_b(\rho)C_b^B}{\sqrt{6}} \\ \frac{m_b(\rho)C_b^R + m_b(\rho)C_b^G + m_b(\rho)C_b^B}{\sqrt{3}} \end{pmatrix} \tag{7}$$

The O_1 and O_2 channels are independent to highlights but are still sensitive to geometry, shading and the intensity of the light source. The O_3 channel is the intensity and contains no invariance properties. We can also take the hue for the opponent colours (*hue*):

$$hue = \arctan\left(\frac{O_1}{O_3}\right) = \arctan\left(\frac{\sqrt{3}(C_b^R - C_b^G)}{C_b^R + C_b^G + C_b^B}\right) \tag{8}$$

The chromatic opponent colours (*chroma*) can also be computed.

$$\begin{pmatrix} C_a \\ C_b \end{pmatrix} = \begin{pmatrix} \frac{O_1}{O_3} \\ \frac{O_2}{O_3} \end{pmatrix} = \begin{pmatrix} \sqrt{\frac{3}{2}}\left(\frac{C_b^R - C_b^G}{C_b^R + C_b^G + C_b^B}\right) \\ \sqrt{\frac{1}{2}}\left(\frac{C_b^R + C_b^G - 2C_b^B}{C_b^R + C_b^G + C_b^B}\right) \end{pmatrix} \tag{9}$$

Both the hue and the chromatic opponent colours are invariant to lighting geometry and specularities [1].

3 Image Representation and Pathology Recognition

For the recognition of pathologies through local information, k-means method shows to be the most robust for the image description. This method is also known as the *Bag of Words*. As a brief description lets suppose we have a set of data points $\{x_n\} \in \mathbb{R}^d$, where d is the dimension of the feature space. K-means consist in searching for K cluster centres, the vectors $\{c_k\} \in \mathbb{R}^d$ and finds an assignment function $A : \{x_n\} \to \{c_k\}$ until the sum of the squares of the distances to each data point to its closest vector c_k is minimum. We estimated an optimal value of 400 clusters. For bigger values the computational cost is not justified.

Learning the Patterns This is performed with a Support Vector Machine (SVM). They rely on the *Support Vectors* to determine a maximum margin hyperplane that separates the classes. Let us assume we have N input vectors $x_1, x_n, ..., x_N$ with the corresponding labels $y_1, y_n, ..., y_N$ where $y_n \in \{-1, 1\}$. The data points are transformed into a higher dimensional space by means of a linear mapping, $\phi(x)$. The maximum-margin hyperplane is defined by $g(x) = w^T \phi(x) + b$, where w is the weight vector and b is a bias parameter. To find the maximum margin is equivalent to find:

$$\min \frac{1}{2}\|w\|^2 + C\sum_{n=1}^{N} \xi_n \tag{10}$$

subject to the constraints:

$$y_n(w^T \phi(x_n) + b) - 1 + \xi_n \geq 0 \quad \forall \quad n \in \{1, ..., N\} \tag{11}$$

where the parameter C controls the trade-off between the slack variable penalty (ξ_n) and the size of the margin. In this work we estimated that, between the *Intersection*, the χ^2 and the *Linear Kernel*, the best results were attained with the *Intersection Kernel*.

4 Results

The image set is composed of endoscopic images from the Barrett's oesophagus with a total of 250 images classified according to Singh's proposal and they were also annotated. From those, 61 are from Type A (normal), and the remaining 189 are from Types B, C and D (pathologic), reducing the problem to a binary classification system. For the training stage we considered 50 normal and 50 pathologic images in order to balance the dataset. To attain the best regularization parameter, we performed a three-fold cross-validation scheme. We began by analysing the performance with the gray descriptor. Each step consist in adding consecutively colour information to this descriptor according to Section 2.2. In order to attain stable results we repeated each experiment 10 times and calculated the mean values of the Mean Average Precision (MAP) and the corresponding standard deviations.

Descriptors	MAP
$\{gray\}$	0.794 ± 0.038
$\{opp\}$	0.748 ± 0.083
$\{gray; rgb\}$	0.819 ± 0.069
$\{gray; rgb; ratio\}$	0.815 ± 0.050
$\{gray; rgb; ratio; log\}$	0.805 ± 0.049
$\{gray; rgb; ratio; log; opp\}$	0.840 ± 0.042
$\{gray; rgb; ratio; log; opp; chroma; hue\}$	0.811 ± 0.049

Table 1: MAP for the tested sets of descriptors.

From Table 1 we observe that the set $\{gray; rgb; ratio; log; opp\}$ gives the highest MAP. For the referred set of descriptors we present the *sensitivity* and *specificity* in Table 2.

Descriptor set	*sensitivity*	*specificity*
$\{gray; rgb; ratio; log; opp\}$	0.788	0.891

Table 2: *Sensitivity* and *Specificity* for the best set of descriptors.

5 Conclusions

The obtained results show that adding information from the RGB channels independently, the ratio of the RGB channels, the logarithmic of these ratios and the information from the opponent colours gives the best results. With this set of local descriptors we attained a MAP of 84.0%, a *sensitivity* of 78.8% and a *specificity* of 89.1%. These high values obtained allow to conclude that the use of local descriptors such as SIFT actually give very good results mainly due to its invariance properties and to the very special conditions that these images are acquired. The addition of colour information improved the classifier performance in about 5% comparing to the gray descriptor. The fact that the opponent colour system highly decorrelates the RGB channels explains this gain in performance.

Acknowledgments

This work was financially supported by FCT (Portuguese Science Foundation) grant PTDC/EIA-CCO/109982/2009. We also would like to thank project PEst-OE/EEI/LA0008/2013.

References

[1] Theo Gevers, Arjan Gijsenij, Joost Van de Weijer, and Jan-Mark Geusebroek. *Color in computer vision: Fundamentals and applications*, volume 24. Wiley. com, 2012.

[2] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[3] Manabu Muto, Takahiro Horimatsu, Yasumasa Ezoe, Shuko Morita, and Shinichi Miyamoto. Improving visualization techniques by narrow band imaging and magnification endoscopy. *Journal of Gastroenterology and Hepatology*, 24(8):1333–1346. ISSN 1440-1746. doi: 10.1111/j.1440-1746.2009.05925.x.

[4] Eric Nowak, Frédéric Jurie, and Bill Triggs. Sampling strategies for bag-of-features image classification. In *Computer Vision–ECCV 2006*, pages 490–503. Springer, 2006.

[5] Steven A Shafer. Using color to separate reflection components. *Color Research & Application*, 10(4):210–218, 1985.

[6] R Singh, GK Anagnostopoulos, K Yao, H Karageorgiou, PJ Fortun, A Shonde, K Garsed, PV Kaye, CJ Hawkey, K Ragunath, et al. Narrow-band imaging with magnification in barrett's esophagus: validation of a simplified grading system of mucosal morphology patterns against histology. *Endoscopy*, 40(6):457–463, 2008.

References

Bashkatov, A., Genina, E., Kochubey, V., & Tuchin, V. (2005). Optical properties of human skin, subcutaneous and mucous tissues in the wavelength range from 400 to 2000 nm. *Journal of Physics D: Applied Physics*, 38(15), 2543.

Bishop, C. M. & Nasrabadi, N. M. (2006). *Pattern recognition and machine learning*, volume 1. springer New York.

Bryan, R. T., Billingham, L. J., & Wallace, D. M. A. (2008). Narrow-band imaging flexible cystoscopy in the detection of recurrent urothelial cancer of the bladder. *BJU International*, 101(6), 702–706.

Coimbra, M. & Cunha, J. (2006). Mpeg-7 visual descriptors 8212; contributions for automated feature extraction in capsule endoscopy. *Circuits and Systems for Video Technology, IEEE Transactions on*, 16(5), 628 – 637.

Coimbra, M., Riaz, F., Areia, M., Silva, F., & Dinis-Ribeiro, M. (2010). Segmentation for classification of gastroenterology images. In *Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE* (pp. 4744 –4747).

Ganz, M., Yang, X., & Slabaugh, G. (2012). Automatic segmentation of polyps in colonoscopic narrow-band imaging data. *Biomedical Engineering, IEEE Transactions on*, 59(8), 2144 –2151.

Gevers, T., Gijsenij, A., Van de Weijer, J., & Geusebroek, J.-M. (2012). *Color in computer vision: Fundamentals and applications*, volume 24. Wiley. com.

Hecht, E. (1998). *Optics*. Addison-Wesley, 4th edition.

Jobe, B. A., Thomas, C. R., & Hunter, J. G. (2009). *Esophageal cancer: principles and practice*. Demos Medical Publishing.

Karargyris, A. & Bourbakis, N. (2009). Identification of polyps in wireless capsule endoscopy videos using log gabor filters. In *Life Science Systems and Applications Workshop, 2009. LiSSA 2009. IEEE/NIH* (pp. 143 –147).

Karkanis, S., Iakovidis, D., Karras, D., & Maroulis, D. (2001). Detection of lesions in endoscopic video using textural descriptors on wavelet domain supported by artificial neural network architectures. In *Image Processing, 2001. Proceedings. 2001 International Conference on*, volume 2 (pp. 833 –836 vol.2).

Karkanis, S., Iakovidis, D., Maroulis, D., Karras, D., & Tzivras, M. (2003). Computer-aided tumor detection in endoscopic video using color wavelet features. *Information Technology in Biomedicine, IEEE Transactions on*, 7(3), 141 –152.

Khademi, A. & Krishnan, S. (2007). Multiresolution analysis and classification of small bowel medical images. In *Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE* (pp. 4524 –4527).

Kwitt, R. & Uhl, A. (2007). Modeling the marginal distributions of complex wavelet coefficient magnitudes for the classification of zoom-endoscopy images. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on* (pp. 1 –8).

Liedlgruber, M. & Uhl, A. (2011). Computer-aided decision support systems for endoscopy in the gastrointestinal tract: A review. *Biomedical Engineering, IEEE Reviews in*, 4, 73 –88.

Lima, C. S., Barbosa, D., Ramos, J., Tavares, A., Monteiro, L., & Carvalho, L. (2008). Classification of endoscopic capsule images by using color wavelet features, higher order statistics and radial basis functions. In *Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE* (pp. 1242 –1245).

Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, volume 2 (pp. 1150–1157).: Ieee.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2), 91–110.

Maji, S., Berg, A. C., & Malik, J. (2008). Classification using intersection kernel support vector machines is efficient. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on* (pp. 1–8).: IEEE.

Muto, M., Horimatsu, T., Ezoe, Y., Morita, S., & Miyamoto, S. (2009). Improving visualization techniques by narrow band imaging and magnification endoscopy. *Journal of Gastroenterology and Hepatology*, 24(8), 1333–1346.

Niemz, M. (2007). *Laser-tissue interactions: fundamentals and applications*. Springer.

Nowak, E., Jurie, F., & Triggs, B. (2006). Sampling strategies for bag-of-features image classification. In *Computer Vision–ECCV 2006* (pp. 490–503). Springer.

Poh, C. K., Htwe, T. M., Li, L., Shen, W., Liu, J., Lim, J. H., Chan, K. L., & Tan, P. C. (2010). Multi-level local feature classification for bleeding detection in wireless capsule endoscopy images. In *Cybernetics and Intelligent Systems (CIS), 2010 IEEE Conference on* (pp. 76 –81).

Shafer, S. A. (1985). Using color to separate reflection components. *Color Research & Application*, 10(4), 210–218.

Shawe-Taylor, J. & Cristianini, N. (2004). *Kernel methods for pattern analysis*. Cambridge university press.

Singh, R., Anagnostopoulos, G., Yao, K., Karageorgiou, H., Fortun, P., Shonde, A., Garsed, K., Kaye, P., Hawkey, C., Ragunath, K., et al. (2008). Narrow-band imaging with magnification in barrett's esophagus: validation of a simplified grading system of mucosal morphology patterns against histology. *Endoscopy*, 40(6), 457–463.

Sousa, A., Dinis-Ribeiro, M., Areia, M., & Coimbra, M. (2009). Identifying cancer regions in vital-stained magnification endoscopy images using adapted color histograms. In *Image Processing (ICIP), 2009 16th IEEE International Conference on* (pp. 681 –684).

Sousa, R., Coimbra, M. T., Ribeiro, M.-D., & Pimentel-Nunes, P. (2013). Impact of svm multiclass decomposition rules for recognition of cancer in gastroenterology images. In *CBMS*. ACCEPTED.

Stehle, T., Auer, R., Gross, S., Behrens, A., Wulff, J., Aach, T., Winograd, R., Trautwein, C., & Tischendorf, J. (2009). Classification of colon polyps in NBI endoscopy using vascularization features. In N. Karssemeijer & M. L. Giger (Eds.), *Medical Imaging 2009: Computer-Aided Diagnosis*, volume 7260 Orlando, USA: SPIE.

Van De Weijer, J., Gevers, T., & Geusebroek, J.-M. (2005). Edge and corner detection by photometric quasi-invariants. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(4), 625–630.

Vedaldi, A. & Fulkerson, B. (2008). VLFeat: An open and portable library of computer vision algorithms. http://www.vlfeat.org/.