

University of Coimbra
Faculty of Sciences and Technology
Department of Physics

DESCRIPTION AND CLASSIFICATION OF
CONFOCAL ENDOMICROSCOPIC IMAGES FOR
THE AUTOMATIC DIAGNOSIS OF THE
INFLAMMATORY BOWEL DISEASE

Sara Queirós Couceiro

Coimbra, 2012

Description and Classification of Confocal Endomicroscopic Images for the Automatic Diagnosis of the Inflammatory Bowel Disease

Advisor: Prof. Dr. João Pedro Barreto
Co-Advisor: Prof. Dr. Pedro Figueiredo

Committee:

Prof. Dr. António Miguel Lino Santos Morgado
Prof. Dr. Bernardete Martins Ribeiro
Prof. Dr. José Pedro Figueiredo
Prof. Dr. João Pedro Barreto

Thesis submitted in partial fulfillment of the requirements
for the Master's Degree in Biomedical Engineering

Department of Physics
Faculty of Sciences and Technology
University of Coimbra

Coimbra, 2012

This copy of the thesis has been supplied on condition that anyone who consults it is understood to recognize that its copyright rests with its author and that no quotation from the thesis and no information derived from it may be published without proper acknowledgment.

Esta cópia da tese é fornecida na condição de que quem a consulta reconhece que os direitos de autor são pertença do autor da tese e que nenhuma citação ou informação obtida a partir dela pode ser publicada sem a referência apropriada.

To Mom and Dad,
for teaching me to grow up
in this beautiful, imperfect world...

Acknowledgements

My first acknowledgement goes to my advisor Prof. Dr. João Pedro Barreto, for introducing me the challenging world of investigation and for constantly spreading positive motivation during this work. I am extremely grateful for his permanent support, guidance and availability for productive discussions.

I would also like to thank to Prof. Dr. Pedro Figueiredo and Dr. Paulo Freire for their collaboration in this research project.

Thanks to Aniana for giving me all the important details of her previous work and for always claryfing my doubts.

To my lab colleagues, I owe the everyday companionship. Thank you for letting me join the group and for sharing ideas, talks and laughs. A special thanks goes to Miguel, for his precious tips and advices and patience for repeated questions and explanations.

I must thank to my friends for making these past academic years an outstanding experience. You turned simple moments into unforgettable memories. Thanks for all the encouragement during this year; it would not have been the same if you had not been part of it.

Last but not least, I want to thank my parents for all the love I was pleased to get from them. Thank you for standing by my side even in the most hard moments and for making me believe that I am able to keep going straight. You never let me give up. No words can express my gratitude for you.

Contents

1	Introduction	1
1.1	Motivation	1
1.2	Confocal Endomicroscopy	2
1.2.1	The Basic Principles of Confocal Microscopy	2
1.2.2	Image Artifacts	3
1.2.3	Overview on Intestinal Histology	3
1.3	Previous Work	4
1.3.1	Database Overview	4
1.4	Problem Definition, Objectives and Contributions	5
1.5	Thesis Overview	7
2	State-of-the-art in Classification and Related Topics	8
2.1	Machine Learning - The Basics	8
2.2	SVM	8
2.2.1	The Statistical Learning Theory	9
2.2.2	Linear SVM: The separable case	9
	Rigid-Margin SVM	11
2.2.3	The non-linearly separable case	12
	Soft-Margin SVM	12
	Non-Linear SVM	14
2.2.4	The Multi-Class Problem	15
2.3	Ensemble Learning - The Random Forest Classifier	16
2.3.1	Brief Introduction to Decision Trees	16
2.3.2	The Random Forest Algorithm	16
2.4	Feature Selection	18
2.4.1	Texture Analysis	18
	Histogram Moments	18
	Gray Level Co-Occurrence Matrix (GLCM)	19
	Gray Level Run-Length Matrix (GLRLM)	20
	Law's Texture Energy Measures	20
2.4.2	SIFT descriptor	21
2.5	Performance Evaluation	22

2.5.1	The Confusion Matrix	22
2.5.2	The ROC Curve	23
2.5.3	The PR Curve	24
2.5.4	Cross-Validation	25
3	Image Segmentation	27
3.1	The segmentation task: The crypts	27
3.2	Centers Detection: A symmetry energy approach	28
3.2.1	Symmetry Energy - Background	28
3.2.2	Symmetry Energy - The Algorithm	29
3.2.3	Symmetry Energy - Tuning parameters	29
3.2.4	Local Maxima of Symmetry Energy	30
3.2.5	Detection Results	31
3.3	Crypts Segmentation	33
3.3.1	Overview	33
3.3.2	Basic Concepts on Conic Curves	33
	The conic equation	33
	Ellipse parameters	33
3.3.3	Brief Introduction to RANSAC	34
3.3.4	The Segmentation Algorithm	34
	Edge Detection	34
	Searching for boundary points	35
	Ellipse Fitting and RANSAC	37
3.3.5	Segmentation Results	38
3.4	SVM Discriminator	39
3.4.1	The Normalization Step	39
	Affine and Scale Invariance	39
3.4.2	Selection of Features	40
	The Radial Gradient Descriptor	40
3.4.3	SVM Discriminator Scores	42
3.5	Global Evaluation of Crypts Detector	45
4	Image Classification	46
4.1	Classification Strategy	46
4.2	Dataset binary subdivision	46
4.2.1	Features used	47
4.2.2	Image Classification Scores	48
5	Conclusions and Future Work	49
	Bibliography	53

List of Figures

1.1	Confocal Microscopy.	2
1.2	Image Artifacts.	3
1.3	Gastrointestinal architecture: longitudinal (a) and transversal (b) histological cuts and transversal section obtained from CEM (c).	4
1.4	Dataset Statistics.	5
2.1	Linear SVM: The optimal hyperplane.	10
2.2	Soft-Margin SVM: introduction of slack variables.	12
2.3	The kernel trick: mapping the input space into the feature space.	14
2.4	Random Forest.	16
2.5	SIFT descriptor.	21
2.6	The ROC space.	24
2.7	Visual comparison between a ROC curve (a) and a PR curve (b).	25
3.1	Crypts variability.	27
3.2	Results of Kovesi's algorithm: original CEM images and their corresponding outputs (<i>phaseSym</i> , <i>orientation</i> and <i>totalEnergy</i>), using the tuned parameters.	30
3.3	Result of crypts' centers detection. The red points correspond to the local maxima of symmetry energy.	31
3.4	Criterion used to evaluate the success of centers detection stage.	31
3.5	Centers detection: PR curve.	32
3.6	Ellipse parameters.	34
3.7	Result of the application of Canny filter over a CEM image.	35
3.8	Generation of a polar image from the cartesian canny result.	36
3.9	Inverse mapping from the polar to the cartesian space. The detected edge points are highlighted in green.	36
3.10	Fitting ellipse computed from the detected edge points, using a RANSAC procedure.	37
3.11	Final segmentation result.	37
3.12	Criteria used to evaluate the performance of the segmentation algorithm.	38
3.13	The normalization step: Affine and Scale Invariance.	40

3.14	The figures show the radial gradient profiles of two different segmented regions: one corresponds to a crypt (first row) and the other one to a false detection (second row). The ellipses computed in the segmentation stage, denoted by the red dashed ellipses in (a) and (e), are enlarged to include a possible boundary. The patches are then normalized to achieve affine and scale invariance, as shown in (b) and (f), and the radial gradient magnitudes are computed from a polar mapping (see (c) and (g)). Finally, the radial gradient profiles, (d) and (h), are obtained by averaging the radial gradient magnitude values across rows of the radial images. Note that, in (d) and (h), the black dashed vertical lines correspond to the red dashed ellipses in (a) and (e).	41
3.15	SVM Discriminator Performance: The graphs show the 10-fold cross-validation averaged ROC curves of the evaluated models. In (a), we compare the several texture features; in (b) we compare the remaining descriptors.	42
3.16	Examples of TP, FP, TN and FN obtained by the classifier trained with our proposed radial gradient profile.	43
3.17	Results obtained before (a) and after (b) the SVM discriminator when using the radial gradient profile and SIFT as features. In (a), the green ellipses correspond to correctly segmented crypts and the red ellipses denote spurious detections and incorrectly segmented crypts. As we can see from these examples, the classifier is able to discard the spurious segmented crypts, due to its good precision. However, since the recall is not as high as the desired, it also discards some true positives. (In fact, some of the green ellipses from (a) are eliminated by the SVM and do not appear in (b).) . . .	44
3.18	Global evaluation of crypts detector performance. The bars show the cumulative results of the three stages of the proposed algorithm: centers detection, crypts segmentation and SVM discriminator.	45
4.1	Features used for image classification. In each image, it is possible to observe the number of detected crypts and the <i>lattice</i> , computed by Delaunay Triangulation.	47
4.2	Image Classifier Performance: The graphs show the 10-fold cross-validation averaged ROC curves of the evaluated features.	48

List of Tables

1.1	Pentax CEM - main specifications.	4
1.2	Inter and Intra-class variability.	6
2.1	Texture metrics derived from Histogram Moments.	19
2.2	Texture metrics derived from GLCM.	19
2.3	Texture metrics derived from GLRLM.	20
2.4	The Confusion Matrix.	22
3.1	Detection Scores.	32
3.2	Segmentation Scores.	38
3.3	Performance metrics of the several models: The table shows the mean values of ACC, SE, SP, PRC and AUC (computed by averaging the results over the 10-fold cross-validation rounds).	42
4.1	Performance metrics of the several features: The table shows the mean values of ACC, SE, SP, PRC and AUC (computed by averaging the results over the 10-fold cross-validation rounds).	48

Acronyms

1A1 One-Against-One

1AA One-Against-All

ACC Accuracy

AUC Area Under Curve

CEM Confocal Endomicroscopy

FN False Negative

FOV Field-of-View

FP False Positive

FPR False Positive Rate

GI gastrointestinal

GLCM Gray Level Co-Occurrence Matrix

GLRLM Gray Level Run-Length Matrix

IBD Inflammatory Bowel Disease

KKT Karush-Kuhn-Tucker

PR Precision-Recall

PRC Precision

RANSAC RANdom SAmples Consensus

RFB Radial Basis Function

ROC Receiver Operating Characteristic

SE Sensitivity

SIFT Scale Invariant Feature Transform

SP Specificity

SVM Support Vector Machine

TN True Negative

TNR True Negative Rate

TP True Positive

TPR True Positive Rate

Abstract

The Inflammatory Bowel Disease refers to a group of idiopathic inflammatory conditions of the gastrointestinal tract, which increase the risks of developing colorectal cancer. Since the diagnosis of these type of disorders in the early stages strongly increases the chances of successful treatment, the surveillance of suspicious lesions becomes extremely important.

Confocal Endomicroscopy is a newly developed diagnosis tool which provides *in vivo* examination of the gastrointestinal histological architecture. Real-time inspection of the tissue allows to avoid the traditional biopsy procedure, providing an immediate diagnosis. In the future, this technique will certainly gain clinical importance in the detection and surveillance of gastrointestinal disorders.

Since this is a very recent diagnosis tool, the taxonomy of the several disease stages in endomicroscopic images is not clearly defined yet. Thus, the analysis of these images is still a challenging task for experts and requires them to pass through a long training period.

We have an expert-annotated database which contains images from five different pathological stages. In the available dataset, it is possible to observe a huge variability of cases, which makes this a complex multi-class classification task. We aim at using computer based tools to help doctors in the interpretation of endomicroscopic images and, thus, provide useful clinical advices for the diagnosis, while still reducing the learning curve of the physicians.

According to experts, the histological appearance of the gastrointestinal tract provides relevant information about the degree of severity of disorders. Based on this feedback, we rely on the tissue's histological properties to make a split between two main subsets of the dataset: low and high probability of pathology.

We start by focusing on image segmentation tools to build a detector of crypts, which are the most evident histological structures. Then, the arrangement of crypts over the tissue is encoded in a feature vector to perform the final classification.

The current results are promising and encouraging toward forward research. Future goals include increasing the performance of the actual classifier and exploring other relevant histological features to distinguish between the remaining classes.

Resumo

A doença inflamatória intestinal refere-se a um conjunto de condições inflamatórias do trato gastrointestinal, que elevam o risco de ocorrência de cancro colo-rectal. O diagnóstico deste tipo de distúrbios em fases iniciais aumenta consideravelmente as hipóteses de sucesso de tratamento, pelo que o acompanhamento de lesões suspeitas é de extrema importância.

A Endomicroscopia Confocal é uma ferramenta de diagnóstico recente que permite efectuar uma análise *in vivo* da arquitectura histológica do trato gastrointestinal. A análise do tecido em tempo real evita a biópsia tradicional e permite efectuar um diagnóstico imediato. No futuro, esta técnica irá certamente ganhar relevância clínica na detecção e vigilância de perturbações gastrointestinais.

Uma vez que este é um método de diagnóstico recente, a taxonomia dos vários estágios da doença em imagens de Endomicroscopia Confocal ainda não se encontra bem definida. Por este motivo, a análise deste tipo de imagens pode tornar-se difícil mesmo para os especialistas, que necessitam de passar por um período de aprendizagem.

Neste momento, existe uma base de dados anotada, que contém imagens pertencentes a cinco estágios diferentes da doença. O conjunto de dados disponível permite observar uma enorme variabilidade de casos, o que conduz a um problema de classificação multi-classes complexo.

De acordo com os especialistas, a observação do aspecto histológico do trato gastrointestinal permite extrair informação relevante acerca do grau de severidade da doença. Desta forma, as propriedades histológicas do tecido são usadas para fazer uma separação entre dois subconjuntos principais dos dados disponíveis: baixa e alta probabilidade de existência de um estado patológico.

Em primeiro lugar, são exploradas técnicas de processamento de imagem para criar um detector de criptas, que são as estruturas histológicas mais facilmente observáveis. Posteriormente, o arranjo destas estruturas no tecido é codificado num vector de características que é usado para a classificação final.

Os resultados actuais são encorajadores para a continuação do trabalho de investigação. Os objectivos futuros incluem o aumento da performance do classificador e a extracção de outras características histológicas relevantes que permitam distinguir entre as restantes classes.

Chapter 1

Introduction

1.1 Motivation

The Inflammatory Bowel Disease (IBD) refers to a group of inflammatory conditions of the gastrointestinal (GI) tract. IBDs are idiopathic disorders, probably caused by an auto-immune response of the body against its own intestinal tissues. In patients with IBD, there is an increased risk of developing colorectal cancer. Although there are other rarer types of inflammatory bowel diseases, Crohn's disease and ulcerative colitis are the two major forms. The main difference between these two conditions is that, while ulcerative colitis only affects the colon, Crohn's disease may involve any part of the GI tract. Since most of the initial symptoms are undervalued by patients, IBDs are frequently noted at severe and chronic stages of illness. However, like in every clinical context, the diagnosis at early stages strongly increases the chances of successful treatment.

According to gastroenterologists, the analysis of the histological appearance of the GI tract provides highly reliable information for the diagnosis. The common practice consists in removing the suspicious areas detected during endoscopy and then sending them to lab analysis. Despite the helpfulness of biopsies, there are several disadvantages related to this procedure. Tissue removal involves potential risks of bleeding and infection, which require subsequent medical treatment. It may also happen that relevant parts of the tissue are not removed. This will lead to a non-representative biopsy and, possibly, to an underestimated diagnosis [1]. Besides, the wait for lab results slows the process, which is another undesirable point.

1.2 Confocal Endomicroscopy

Confocal Endomicroscopy (CEM) is a recently developed technique that allows the *in vivo* examination of the intestinal mucosa during ongoing endoscopy [2]. The basis behind this new diagnosis tool is the integration of a mini confocal microscope with the distal tip of a conventional endoscope [3].

The main advantage of CEM over the traditional biopsy procedure is that it provides real-time histological examination, allowing an immediate diagnosis. Besides, since no biopsy is needed, the risks associated with tissue removal are avoided.

In the future, CEM will certainly gain importance in clinical gastroenterology [3] and, particularly, in the detection and surveillance of gastrointestinal neoplasia.

1.2.1 The Basic Principles of Confocal Microscopy

Confocal Microscopy, whose principle was patented by Marvin Minsky in 1957, allows to overcome some of the limitations of conventional wide-field fluorescence microscopes.

As shown in Fig. 1.1, a low-powered laser is passed through an illumination pinhole and focused into one single point. Since the same lens is used both as the objective and the condenser, the point of illumination and the point of detection within the examined sample are coincident. The fluorescent light from this spot is passed through a pinhole to a photodetection device and the light that comes from outside of it is rejected. Therefore, at every instance, only one point of the fluorescent specimen is lighted.

In regular microscopy, the whole specimen is illuminated at the same time, which means that the out-of-focus light emitted by the sample also contributes to the image formation. Confocal Microscopy solves this problem by rejecting the light from other focal points, improving the resolution and quality of the collected images.

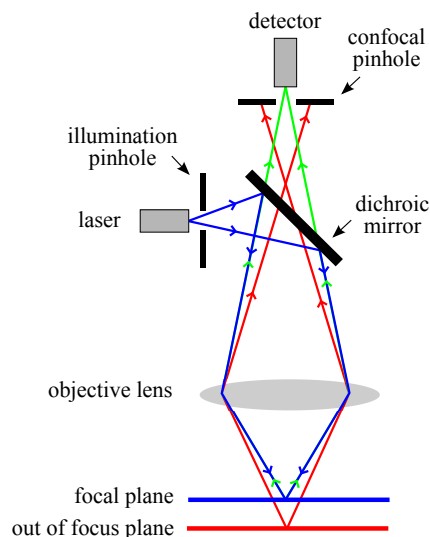


Figure 1.1: Confocal Microscopy.

The detected light is then converted into an electrical signal and recorded by a computer. A gray-scale image is created by scanning the whole sample. At each image pixel, the brightness value corresponds to the intensity of the detected light at that point of the fluorescent specimen. This image is an optical section that represents one focal plane of the observed sample.

This technique depends on the fluorescence of the examined specimen. The most commonly used contrast agents are intravenous fluorescein and topical acriflavine [1].

1.2.2 Image Artifacts

A consistent analysis of the histological architecture is highly dependent on the quality of CEM images. Image acquisition is not a simple procedure for the operator since it requires some practice in endoscope manipulation, implying a previous learning process. This is extremely important to avoid the introduction of image artifacts and the resultant distortion of information.

The endoscope should be stably positioned to avoid motion artifacts (Fig.1.2(a)), which produce a kind of a blurring effect on the images. Besides, keeping the endoscope oriented perpendicularly to the tissue surface avoids slant artifacts (Fig.1.2(b)), which could lead to a misinterpretation of the shape and orientation of histological structures. Another possible image artifact is related to the presence of mucus, fecal content and air bubbles (Fig.1.2(c)) in the bowel lumen.

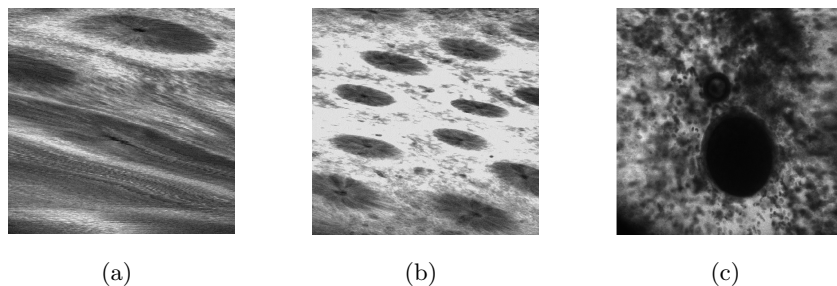


Figure 1.2: Image Artifacts.

1.2.3 Overview on Intestinal Histology

Because this is a very recent technique, the taxonomy of pathological stages in CEM images is not yet clearly defined. Physicians are still exploring the complexity of these images and, even for them, their interpretation might be a challenging task. In what concerns the GI histological architecture, some relevant structures should be paid attention.

The most evident histological structures are the intestinal crypts (also known as crypts of Lieberkühn or intestinal glands), which are tubular invaginations of the epithelium mainly involved in secretion. They also contain stem cells which are responsible for the

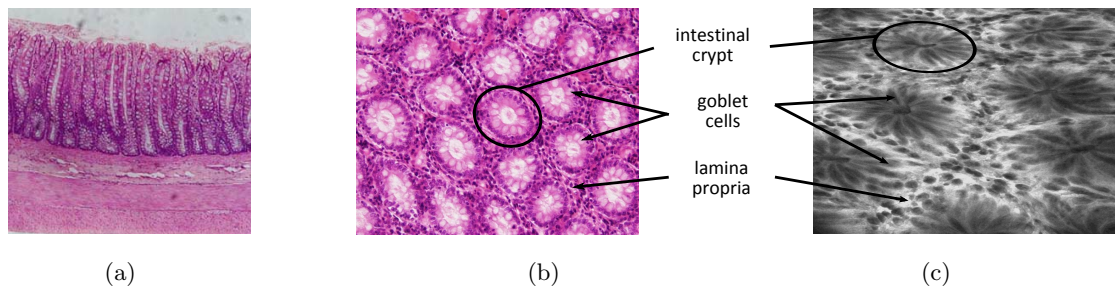


Figure 1.3: Gastrointestinal architecture: longitudinal (a) and transversal (b) histological cuts and transversal section obtained from CEM (c).

ongoing renewal of the epithelial tissue. Uncontrolled proliferation in crypts may lead to colorectal cancer. Goblet cells, whose name comes from their goblet shape, are glandular epithelial cells that are commonly prevalent in crypts. The space between the crypts corresponds to the lamina propria, which is composed of connective tissue.

1.3 Previous Work

The previous work consisted on the establishment of an acquisition protocol for collecting and labeling CEM images in order to create an annotated database [4]. The importance of building a large scale database is related to the need of gathering useful clinical information to be used as a high-quality research tool. This is extremely helpful to provide medical education by increasing the knowledge of the GI architecture in CEM images.

The developed software application allows the doctors to label the collected images and also to interactively store relevant clinical annotations.

1.3.1 Database Overview

The available dataset was collected at the Gastroenterology Services of the University Hospital of Coimbra. The patients were examined by experts using fluorescein as contrast agent and a Pentax CEM device, whose main specifications are shown in the table below.

Table 1.1: Pentax CEM - main specifications.

Excitation Wavelength	488 <i>nm</i>
Range of depth	0 – 250 μm
Resolution	1024 \times 512 or 1024 \times 1024
Range of laser power	0 – 1000 μW
Field-of-View (FOV)	475 \times 475 μm
Frame Rate	1.2 frames per second (1024 \times 512) or 0.7 frames per second (1024 \times 1024)

Images were classified by experts into three main stages: *healthy tissue*, *inflammation* and *neoplasia*. In inflammation and neoplasia stages, it is still possible to identify two different levels of illness severity, which leads to a five class division of the data: 1. *healthy tissue*, 2. *light inflammation*, 3. *severe inflammation*, 4. *low probability of neoplasia* and 5. *high probability of neoplasia*.

At the moment, the database contains information about 18 patients and 192 images with a resolution of 1024×512 pixels. The images are distributed over the five classes as shown in the following graph:

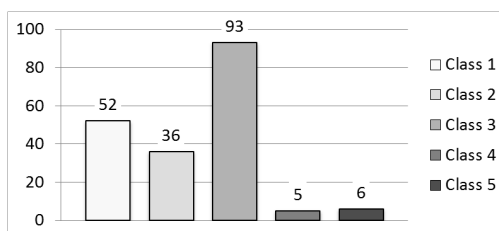


Figure 1.4: Dataset Statistics.

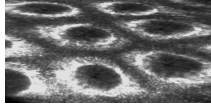
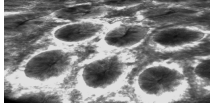
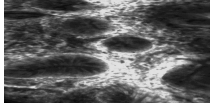
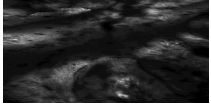
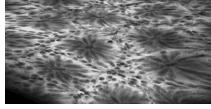
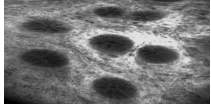
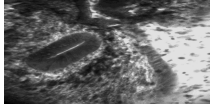
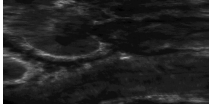
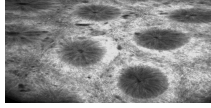
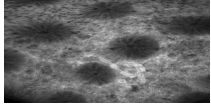
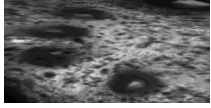
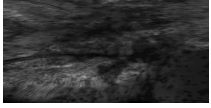
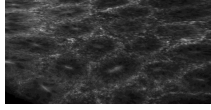
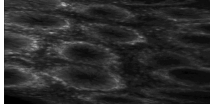
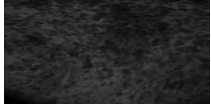
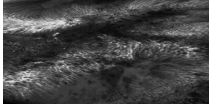
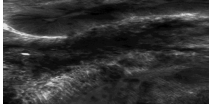
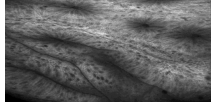
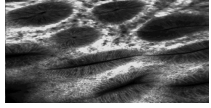
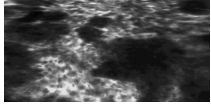
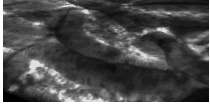
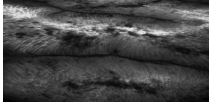
1.4 Problem Definition, Objectives and Contributions

Computer-aided diagnosis tools allow to assist physicians in clinical procedures and have been gaining importance in medical research fields. Due to the difficulty in establishing a clear taxonomy in CEM images, computer based tools can be extremely beneficial to reduce the learning curve of the physicians and to improve the diagnosis accuracy.

In this work, we aim at developing segmentation procedures to reliably highlight relevant histological structures in the tissue in order to aid physicians in image interpretation and provide useful guidances for the diagnosis. The detection of abnormalities in the tissue is helpful to quantify the progression of the disease and assist the surveillance of suspicious lesions. The ultimate goal is to build an automatic classification system to distinguish the several stages of IBD during the endoscopic procedure and help doctors in the classification of pathologies.

The main challenge in this classification problem is related to the intra and inter class variability. As it is shown in Table 1.2, besides the huge variability between images from the same class, the differences between the several classes are not clearly evident. For example, in pairs 1c),2c), and 1d),2d), the differences between the images are quite difficult to detect. Similarly, there doesn't seem to exist notorious differences between classes 4 and 5. In images 2a) and 1e), the observed slant artifacts may mislead the diagnosis. Even trained experts may be tricked by this great visual variability.

Table 1.2: Inter and Intra-class variability.

healthy tissue	tissue with inflammation		tissue with neoplasia		
	light inflammation	severe inflammation	low probability of neoplasia	high probability of neoplasia	
					a)
					b)
					c)
					d)
					e)
1	2	3	4	5	

Due the novelty of the CEM technique, the classification of these images has not been very explored yet. Relevant references can be found in [5,6], where André *et al* propose a content-based retrieval approach to perform classification of endomicroscopic images from an expert-annotated database. To deal with the small FOV of the endomicroscopy, the video sequence is used to perform mosaicking. The mosaics are described using bi-scale dense Scale Invariant Feature Transform (SIFT) features [7], that are quantized into visual words using k-means clustering. The classification is then performed by querying an image database with a histogram of visual words and retrieving the most similar images from the database.

This approach does not take into account the specificities of CEM images and does not produce relevant side results like detection and segmentation of histological structures. Besides, it requires large amounts of labeled data and it is not compact in terms of storage. Therefore, it can be time consuming for large databases producing an on-the-fly implementation.

We propose an alternative solution that tries to explore the CEM images specificities to overcome the above mentioned issues. The presented approach passes by detect and segment crypts as being the most relevant and informative image structures. This is accomplished by using symmetry analysis to find candidates, apply edge detection and ellipse fitting to detect contours and employ an SVM to discard false positives. The detected crypts are then used to represent the image through a specifically developed descriptor that is fed into an SVM classifier for the final diagnosis.

The final classification scheme is substantially simpler than the one proposed by Andre's and, thus, potentially more suitable for an on-the-fly implementation. Besides, crypts detection and segmentation may be employed for other purposes, like assisted image annotation.

During this year, a paper was submitted and accepted in the 3rd International Workshop on Machine Learning in Medical Imaging, 15th International Conference on Medical Image Computing and Computer Assisted Intervention (MLMI MICCAI) 2012.

1.5 Thesis Overview

This thesis is divided in five chapters. The present chapter includes the motivation and objectives of the developed work, as well as a brief introduction to the IBD, the CEM technique and an overview on the GI histology.

In the second chapter, we describe some state-of-the-art classification techniques, giving special attention to Support Vector Machine (SVM), and introduce possible approaches to deal with multi-class tasks. We also refer to some commonly used features, as texture statistics and SIFT, and to the performance metrics used to evaluate a classifier model.

The third chapter describes the image segmentation algorithm developed to identify relevant structures in CEM images. We explain in detail the stages of the proposed algorithm.

In the fourth chapter, we refer to the strategy used to deal with the dataset classification problem. We explain the clinical motivation for the proposed data division and the features used to perform such split.

Finally, in the fifth chapter, we present the conclusions of the developed work and propose future improvements.

Chapter 2

State-of-the-art in Classification and Related Topics

2.1 Machine Learning - The Basics

Machine Learning is a core subfield of Artificial Intelligence that focuses on the development of automatic algorithms that enable a machine to complete tasks, such as recognition, diagnosis, planning and prediction.

The several developed Machine Learning techniques may be split into two main categories: *supervised* and *unsupervised* algorithms. Supervised learning algorithms aim to infer a function from the generalization of a set of correctly labeled examples, so that it is possible to predict the labels of new input samples. In unsupervised learning algorithms, by contrast, the examples provided to the learner are unlabeled, so there is no explicit target. The purpose is to find hidden similarities between the data so that new inputs will be classified by comparison.

One important requirement of Machine Learning algorithms is that they should be robust enough to deal with data noise and outliers, which are common in many datasets.

Another key concept is the generalization power, defined as the ability of the classifier to correctly predict the labels of new data. *Overfitting* occurs if the predictor is too specialized on the training data, having a low success rate when classifying new samples. The classifier will also have a low success rate if the available training data is poorly representative or if the inferred model is too simple. This situation is called *underfitting*.

2.2 SVM

A SVM is a very popular supervised learning technique among machine learning algorithms. It is used in both classification and regression problems and it has a wide range of applications. SVM algorithms have gained importance over the years due to its robustness, high accuracy and effectiveness [8]. Compared to other competing classification methods, they often perform better in terms of generalization performance [9].

2.2.1 The Statistical Learning Theory

SVM formulation is based on the Statistical Learning Theory, which aims to deal with the problem of gaining knowledge from an available set of data. It provides a framework to study how to make inferences, predictions and decisions, and how to construct models from the data.

Let X and Y be the input and the output spaces. In the following discussion, we will consider binary classification and use labels $Y = \{-1, +1\}$ for the two classes, with -1 being a negative example and $+1$ a positive example.

Let $(\mathbf{x}, y) \in (X, Y)$ be a training set of dimension l that is sampled according to some unknown distribution $P(\mathbf{x}, y)$. The main goal of a classifier is to find a mapping $f : X \rightarrow Y$, while minimizing the expectation of the test error, also called the *expected risk* or just *risk*, which is given by:

$$R(f) = \int c(f(\mathbf{x}), y) dP(\mathbf{x}, y) . \quad (2.1)$$

In this expression, $c(f(\mathbf{x}), y)$ is called the loss function and is a measure of the test error, that relates the predicted value of \mathbf{x} with its actual value y . A common loss function used in classification problems is:

$$c(f(\mathbf{x}), y) = \frac{1}{2} |y - f(\mathbf{x})| . \quad (2.2)$$

This function returns 0 if \mathbf{x} is correctly classified and 1 otherwise.

Since $P(\mathbf{x}, y)$ is an unknown distribution, usually it is not possible to minimize the *expected risk*. One way of getting over this is to use the Empirical Risk Minimization induction principle, which replaces the *expected risk* by the *empirical risk*. The *empirical risk* is defined as the mean error rate on the training set and is given by:

$$R_{emp}(f) = \frac{1}{l} \sum_{i=1}^l c(f(\mathbf{x}_i), y_i) . \quad (2.3)$$

2.2.2 Linear SVM: The separable case

Let's assume that there is a hyperplane which separates positive points, $\mathbf{x} \in \{+1\}$, from negative points, $\mathbf{x} \in \{-1\}$. The points lying in this hyperplane satisfy:

$$\mathbf{w}^T \mathbf{x} + b = 0 , \quad (2.4)$$

where \mathbf{x} is a vector point and \mathbf{w} is the weight vector, which is perpendicular to the hyperplane and has norm $\|\mathbf{w}\|$. $\frac{|b|}{\|\mathbf{w}\|}$ represents the perpendicular distance from the origin to the hyperplane.

The separating hyperplane divides the data space into two distinct regions, each one corresponding to one of the classes. In each region, the data points which are closest to the hyperplane are called *support vectors*.

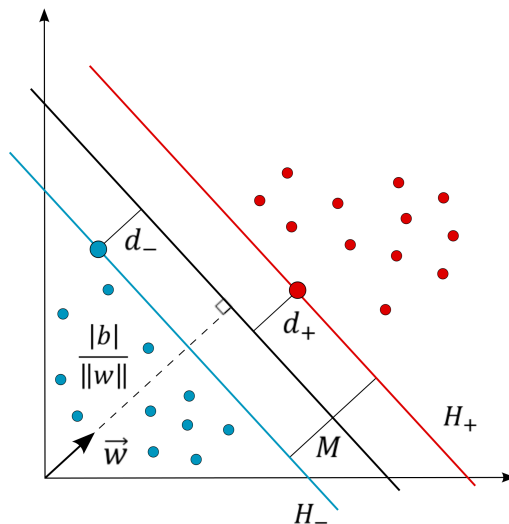


Figure 2.1: Linear SVM: The optimal hyperplane.

Support vectors are considered to be the most important data from the training set, since they are the only data points used to determine the equation of the separating hyperplane.

Let d_+ and d_- be, respectively, the perpendicular distances from the separating hyperplane to the closest positive and negative support vectors. H_+ and H_- are the hyperplanes which are parallel to the separating hyperplane and contain the support vectors. These hyperplanes are defined by:

$$H_+ : \quad \mathbf{w}^T \mathbf{x} + b = +1 \quad (2.5)$$

$$H_- : \quad \mathbf{w}^T \mathbf{x} + b = -1 . \quad (2.6)$$

Note that any point from the training set falls between these two hyperplanes. Thus, every training data satisfy:

$$\mathbf{w}^T \mathbf{x}_i + b \geq +1, \quad y_i = +1 \quad (2.7)$$

$$\mathbf{w}^T \mathbf{x}_i + b \leq -1, \quad y_i = -1 . \quad (2.8)$$

Combining these two inequations yields:

$$y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1, \quad \forall i . \quad (2.9)$$

The distances from H_+ and H_- to the origin are, respectively, $\frac{b+1}{\|\mathbf{w}\|}$ and $\frac{b-1}{\|\mathbf{w}\|}$. The *margin* M is defined as the distance between H_+ and H_- , that is:

$$M = \frac{b+1}{\|\mathbf{w}\|} - \frac{b-1}{\|\mathbf{w}\|} = \frac{2}{\|\mathbf{w}\|} . \quad (2.10)$$

The optimal hyperplane allows to separate data with the maximum margin possible and is determined by minimizing $\|\mathbf{w}\|^2$, subject to constraints. This leads to a quadratic optimization problem.

Rigid-Margin SVM

Rigid Margin SVM defines a rigid decision boundary and does not allow any data point to lie inside the margin. The optimization problem becomes:

Minimize:

$$\frac{1}{2} \|\mathbf{w}\|^2 \quad (2.11)$$

subject to:

$$y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1, \quad \forall i. \quad (2.12)$$

This quadratic optimization problem is solved by switching to an unconstrained Lagrangian formulation [10, 11]. The introduction of positive Lagrangian multipliers α_i , $i = 1, \dots, l$ yields:

$$L_P = \frac{1}{2} \|\mathbf{w}\|^2 - \sum_i \alpha_i [y_i(\mathbf{w}^T \mathbf{x}_i + b) - 1]. \quad (2.13)$$

L_P must be minimized with respect to \mathbf{w}, b and maximized with respect to α_i . The solution is given by the saddle point [10]. This is a convex quadratic optimization problem, since the objective function is itself convex and the the points satisfying the constraints also form a convex set. For this reason, it is possible to make use of the Karush-Kuhn-Tucker (KKT) conditions to solve the problem [10, 11] and, therefore, the gradient of L_P should vanish:

$$\frac{\partial L_P}{\partial \mathbf{w}} = 0 \Rightarrow \mathbf{w} = \sum_i \alpha_i y_i \mathbf{x}_i \quad (2.14)$$

$$\frac{\partial L_P}{\partial b} = 0 \Rightarrow \sum_i \alpha_i y_i = 0. \quad (2.15)$$

Replacing these results in equation 2.13 gives:

Maximize:

$$L_D = \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \mathbf{x}_i \mathbf{x}_j \quad (2.16)$$

subject to:

$$\sum_i \alpha_i y_i = 0 \quad (2.17)$$

$$\alpha_i \geq 1, \quad \forall i. \quad (2.18)$$

This formulation is called the dual problem, while 2.13 represents the primal formulation. The most important reason for using the dual formulation is that the data appear

as dot products between vectors. This property becomes extremely important as it will allow to generalize the problem to deal with non-linearly separable data.

Once the optimization problem is solved for α_i , \mathbf{w} may be determined from 2.14. The parameter b is then found from the KKT condition [11]:

$$\alpha_i(y_i(\mathbf{w}^T \mathbf{x}_i + b) - 1) = 0, \quad \forall i. \quad (2.19)$$

2.2.3 The non-linearly separable case

In many real problems, it is not possible to find a decision boundary that exactly separates the data into two classes. This may happen due to the presence of noise or outliers or even because of the non-linear nature of the problem.

Soft-Margin SVM

One possible approach to deal with non-linearly separable data is to smooth the classifier boundaries, allowing some of the data to lie inside the margin. The hyperplane constraints are relaxed by introducing positive slack variables $\xi_i, i = 1, \dots, l$. Thus:

$$y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad \forall i. \quad (2.20)$$

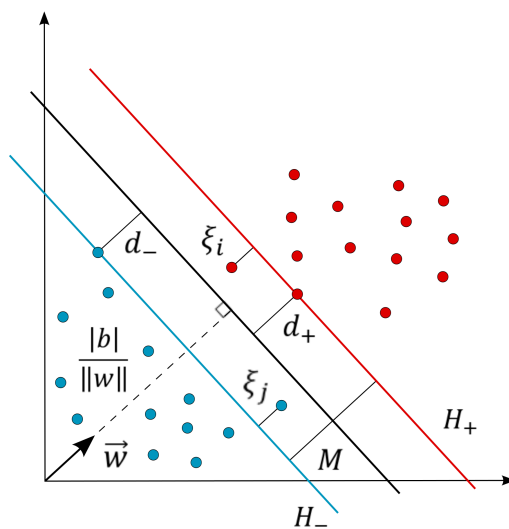


Figure 2.2: Soft-Margin SVM: introduction of slack variables.

If $0 < \xi_i < 1$, then \mathbf{x}_i is well classified, although it is inside the margin. However, if $\xi_i \geq 1$, \mathbf{x}_i is misclassified.

The upper limit of the number of training errors is $\sum_i \xi_i$. Adding this term to the objective function yields:

$$\frac{1}{2} \|\mathbf{w}\|^2 + C \sum_i \xi_i. \quad (2.21)$$

In this expression, C is called the *regularization parameter* and represents a penalty factor to the training errors.

The optimization problem is:

Minimize:

$$\frac{1}{2} \|\mathbf{w}\|^2 + C \sum_i \xi_i \quad (2.22)$$

subject to:

$$y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad \forall i. \quad (2.23)$$

Just like in Rigid Margin SVM, the problem is solved by switching to the Lagrangian formulation:

$$L_P = \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_i \xi_i - \sum_i \alpha_i [y_i(\mathbf{w}^T \mathbf{x}_i + b) - 1 + \xi_i] - \sum_i \mu_i \xi_i. \quad (2.24)$$

Again, according to the KKT conditions, the derivatives of L_P are set to zero:

$$\frac{\partial L_P}{\partial \mathbf{w}} = 0 \Rightarrow \mathbf{w} = \sum_i \alpha_i y_i \mathbf{x}_i \quad (2.25)$$

$$\frac{\partial L_P}{\partial b} = 0 \Rightarrow \sum_i \alpha_i y_i = 0 \quad (2.26)$$

$$\frac{\partial L_P}{\partial \xi_i} = 0 \Rightarrow C - \alpha_i - \mu_i = 0. \quad (2.27)$$

Replacing these results in 2.24 leads to the dual formulation of the problem:

Minimize:

$$L_D = \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \mathbf{x}_i \mathbf{x}_j \quad (2.28)$$

subject to:

$$\sum_i \alpha_i y_i = 0 \quad (2.29)$$

$$0 \leq \alpha_i \leq C, \quad \forall i. \quad (2.30)$$

As before, once α_i is determined, \mathbf{w} and b are found, respectively, from 2.25 and from the KKT condition [11]:

$$\alpha_i (y_i(\mathbf{w}^T \mathbf{x}_i + b) - 1 + \xi_i) = 0, \quad \forall i. \quad (2.31)$$

Non-Linear SVM

In Soft-Margin SVM, the decision boundary is a linear function of the data. Non-Linear SVM deals with non-linearly separable data using an approach that does not try to fit the problem into a linear model.

The main idea of Non-Linear SVM is to apply a suitable non-linear transformation to map the problem to a new space, called the *feature space*, where a linear model can be used. The linear model in the feature space corresponds to a non-linear model in the input space. This procedure is known as the *Kernel Trick*.

The reason for doing this is based on Cover's Theorem. According to this theorem, an input space with non-linearly separable data can be mapped into a higher dimensional space (possibly infinite dimensional), in which data has high probability of being linearly separable. This will be true as long as the mapping transformation is non-linear and the dimension of the feature space is high enough.

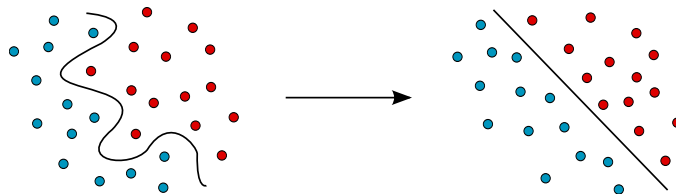


Figure 2.3: The kernel trick: mapping the input space into the feature space.

Let Φ be the function which maps the input space \mathcal{T} (low dimensional) into the feature space \mathcal{H} (high dimensional). Thus:

$$\Phi : \mathcal{T} \rightarrow \mathcal{H} . \quad (2.32)$$

It is not hard to understand that, if the dimension of the feature space is too high, the computation of the mapping function may become too complex to be done in practice. Thus, in these cases, it is not viable to work with Φ explicitly.

However, it is not really necessary to know Φ explicitly, since the only information that is required in the feature space is the dot product between the data.

Given this, we use a kernel function \mathcal{K} , which receives two data points in the input space and returns their dot product in the feature space:

$$\mathcal{K}(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j) . \quad (2.33)$$

Since the kernel function uses an implicit mapping, we can directly apply it in the input space instead of calculating $\Phi(x_i)$ and $\Phi(x_j)$ and then taking the dot product. This makes the procedure much simpler.

The most used kernels are the following:

- polynomial of degree q :

$$(\mathbf{x}_i \mathbf{x}_j + 1)^q ; \quad (2.34)$$

- Radial Basis Function (RFB):

$$\exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma}\right); \quad (2.35)$$

- sigmoid function:

$$\tanh(\kappa\mathbf{x}_i\mathbf{x}_j - \delta). \quad (2.36)$$

The choice of the kernel strongly affects the success of an SVM classifier. Although there aren't clear rules for selecting the most effective kernel for a particular classification task, the RBF function usually offers good performance [12] and it is definitely the most popular kernel choice.

2.2.4 The Multi-Class Problem

There are several available algorithms to efficiently deal with the binary classification problem. In fact, most of the existing techniques have been designed for the two class case, so the multi-class framework requires a little more attention.

Although SVM algorithms were initially created to solve binary tasks, they may be changed to handle multi-class problems. Some of the proposed SVM extensions to the multi-class case involve the addition of extra parameters and constraints to the optimization problem [13]. These formulations may result in a high computational effort if the number of classes is too large.

Another possible approach consists in decomposing the multi-class task into several binary problems that may be solved using binary classifiers. The most common strategies for such decomposition are One-Against-All (1AA) and One-Against-One (1A1) [8].

One-Against-All

This is the simplest approach, which consists in dividing a K classes problem into K binary problems. Thus, for K classes, it is required to build K binary classifiers, each of these separating one class from all the other $K - 1$ classes. The training set of the k^{th} classifier is made of positive examples belonging to class k and negative examples belonging to the other $K - 1$ classes. The class label of a new input sample is determined by the maximum output of the K classifiers.

One-Against-One

This technique compares each pair of classes, meaning that $\frac{K(K-1)}{2}$ binary classifiers are needed. When testing an unknown sample, each binary classifier gives one vote and the overall classification is determined by the winning class.

Although some comparative studies of the performance of 1A1 and 1AA have already been published [8, 14, 15], there are no straight guidelines to clearly choose between these two techniques.

2.3 Ensemble Learning - The Random Forest Classifier

Ensemble learning algorithms have recently turned out to be very popular due to their high accuracy. These methods classify instances by aggregating the results of several classifiers [16]. The idea is to combine multiple *weak* learners in order to obtain a *stronger* learner and yet increase predictive performance.

2.3.1 Brief Introduction to Decision Trees

Decision trees are one of the most widespread used learning tools. They use tree-like structured models to deal with both regression and classification problems, based on understandable rules that can be readily stated.

The tree structure always begins with a root node that corresponds to the whole feature space. Each internal node contains a binary test that splits the data into two disjoint regions. The leaf nodes designate the target classes.

This arrangement naturally leads to a classification process that is based on recursive partitioning. The input data travels top-down in the tree until no further subdivision is possible, that is, until a leaf node is reached.

2.3.2 The Random Forest Algorithm

Like stated above, Random Forests are ensembles of decision trees. The crucial aspect about these trees is that they are grown from a set of independent and identically distributed random vectors [17], meaning that some kind of randomness is injected in the training procedure.

After generating a large number of trees in the ensemble, the classification of a testing sample is found by majority voting [16–18]. The sample travels each one of the trees in the forest until it reaches a leaf, being assigned a set of leaf indexes. The overall classification is then determined by combining all the unit votes.

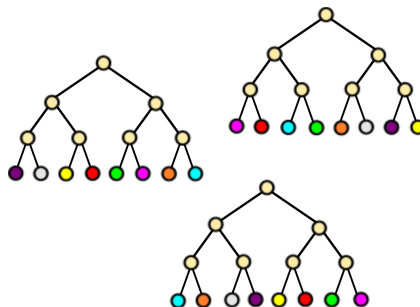


Figure 2.4: Random Forest.

Random Forest algorithm considers two levels of randomness. Actually, randomness is inserted at two distinct points: in the sub-sampling of the training data and in the selection of the binary test for each node [18].

1st randomization level

One of the most known approaches for training data randomization is bagging [19], which stands for *bootstrap aggregation*.

Given a set of label data D , this method repeatedly selects a random subsample, with replacement. Therefore, a sequence of different training subsets is generated, each one consisting of independent samples with the same distribution as D . This procedure allows each one of the trees of the ensemble to grow from a different training subset.

2nd randomization level

At each internal node, the binary test is determined as the one that best separates the data. If bagging was used alone, the best splitter would be chosen among all the available binary tests, which would be computationally very expensive. The injection of randomness at this point of the algorithm performs a random pre-selection of m_{try} binary tests. The best splitter is then picked from this subset.

This additional randomization step strongly reduces computational effort of the growth of trees, speeding up the training process.

Given this, the Random Forest algorithm may be summarized as follows [16]:

- Let n_{trees} be the total number of trees in the ensemble. For each tree:
 - Select a bootstrap training subset from the whole available training data.
 - Grow an unpruned decision tree from this training subset.
 - At each internal node, choose the best splitter through a partially random selection, as explained above.
- Classify new instances by aggregating the unit votes of the n_{trees} trained trees.

Compared to other current classification algorithms, Random Forests are highly accurate predictors [20]. When applied to decision trees, the ensemble learning technique improves the overall accuracy while maintaining the advantages of individual classifiers [20].

Apart from the high accuracy, the main advantage of Random Forests over traditional classifiers is the speed, since they are much faster in training and testing [18]. Moreover, Random Forests also perform efficiently when dealing with large datasets and, according to Breiman [17], they resist to *overfitting*.

2.4 Feature Selection

The success of a classifier strongly depends on the features used for training and testing the model. For each particular classification task, different features may be relevant, depending on the nature of the problem and on the classes we want to distinguish. For this reason, the selection of the appropriate features is a crucial step to achieve a high classification performance.

Here we briefly present some features broadly used in image classification tasks.

2.4.1 Texture Analysis

In general, textures refer to regular repetitions of elements or patterns. They consist of groups of mutually related pixels, called texture primitives or texture elements (*texels*). A primitive is defined to be a set of adjacent pixels in the same direction and with the same gray level.

Texture analysis is used in a range of studies for texture description and classification, image segmentation, shape identification and object recognition. It is not possible to select one single technique to describe the wide variety of existing textures. In fact, there are many texture representation methods used to characterize these complex image visual patterns.

There are two main approaches to this issue, which refer to syntactical and statistical methods. Syntactic methods are based on the arrangement of primitives, defining a texture region as a set of predefined primitives and construction rules. Statistical methods describe the spatial distribution of gray levels by deriving a set of statistics from local features.

While syntactic techniques work well when analyzing artificial textures, the statistical approach is more used in natural textures, which are composed of patterns of irregular sub-elements. For this reason, syntactical approach will not be explored here. The presented analysis techniques refer to statistical methodologies.

Histogram Moments

A histogram is a simple representation of the gray level distribution in an image.

Let L be the number of distinct gray levels z of an input image and $p(z_i)$ the corresponding image histogram. From this, it is possible to compute a set of statistical metrics which characterize the image texture.

Note that these texture measures are related to first-order statistics, since they only consider the original image values and exclude pixel neighborhood relationships.

Table 2.1: Texture metrics derived from Histogram Moments.

Mean (first moment)	$m = \sum_{i=0}^{L-1} z_i p(z_i)$
Variance (second moment)	$\sigma^2(z) = \sum_{i=0}^{L-1} (z_i - m)^2 p(z_i)$
Skewness (third moment)	$\mu_3(z) = \sum_{i=0}^{L-1} (z_i - m)^3 p(z_i)$
Kurtosis (fourth moment)	$\mu_4(z) = \sum_{i=0}^{L-1} (z_i - m)^4 p(z_i)$
Energy or Uniformity	$U = \sum_{i=0}^{L-1} p^2(z_i)$

Gray Level Co-Occurrence Matrix (GLCM)

This method is an estimation of the joint probabilities of pairs of gray level intensities [21].

Consider a position operator P defined in terms of a distance d and an angle θ . GLCM is a square matrix, $C_{k \times k}$, where k is the number of distinct gray levels in the input image. Each element c_{ij} is the number of times that pixels with gray level z_i occur relative to pixels with gray level z_j , in the position specified by P , with $1 \leq i, j \leq k$. C is normalized by the number of point pairs satisfying P .

Numeric features computed from GLCM can be used to represent and compare textures. These features are related to second-order statistics, since they consider the relative position of pixels to each other.

Table 2.2: Texture metrics derived from GLCM.

Variance	$\sum_i \sum_j (i - j)^2 c_{ij}$
Energy	$\sum_i \sum_j c_{ij}^2$
Entropy	$-\sum_i \sum_j c_{ij} \log_2 c_{ij}$
Homogeneity	$\sum_i \sum_j \frac{c_{ij}}{1 + i - j }$
Correlation	$\sum_i \sum_j \frac{(i - \mu)(j - \mu)c_{ij}}{\sigma^2}$

Gray Level Run-Length Matrix (GLRLM)

As explained above, each primitive is described by its gray level, length and direction.

Given an input image $I_{N \times N}$, each element of the GLRLM, $p(i, j)$, represents the number of primitives of all directions with pixels of gray level i and run-length j [22].

Many texture features may be derived from GLRLM. Let L be the number of distinct gray levels, N the maximum run length, n_r the number of total runs and n_p the number of pixels in the input image. The following table contains the five traditional statistics.

Table 2.3: Texture metrics derived from GLRLM.

Short Run Emphasis	$\frac{1}{n_r} \sum_{i=1}^L \sum_{j=1}^N \frac{p(i, j)}{j^2}$
Long Run Emphasis	$\frac{1}{n_r} \sum_{i=1}^L \sum_{j=1}^N p(i, j) \cdot j^2$
Gray Level Nonuniformity	$\frac{1}{n_r} \sum_{i=1}^L \left(\sum_{j=1}^N p(i, j) \right)^2$
Run Len Nonuniformity	$\frac{1}{n_r} \sum_{j=1}^N \left(\sum_{i=1}^L p(i, j) \right)^2$
Run Percentage	$\frac{n_r}{n_p}$

Law's Texture Energy Measures

Laws identified a set of properties for describing texture, which are determined by evaluating Average Gray Level, Edges, Spots, Ripples and Waves [23].

$$\begin{aligned}
 \text{Level} & L_5 = [1, 4, 6, 4, 1] \\
 \text{Edges} & E_5 = [-1, -2, 0, 2, 1] \\
 \text{Spots} & S_5 = [-1, 0, 2, 0, 1] \\
 \text{Ripples} & R_5 = [1, -4, 6, -4, 1] \\
 \text{Waves} & W_5 = [-1, 2, 0, -2, -1]
 \end{aligned} \tag{2.37}$$

These vectors are mutually multiplied to generate a total of 25 masks of dimension 5×5 . The convolution of these masks with the input image results in a set of images, from which energy statistics may be computed.

2.4.2 SIFT descriptor

The SIFT algorithm, published by David Lowe, is a widely used approach for the detection of interest points and the extraction of local features in an image.

Because the produced descriptors are highly distinctive [7], SIFT became one of the most popular choices for feature description of image keypoints. It is particularly important in classification and object recognition.

For each keypoint, the SIFT descriptor is computed from the local image gradients. The gradient magnitudes and orientations are computed over a 16×16 neighborhood around the keypoint location. This 16×16 patch is then divided into 4×4 subregions and, for each one of them, an orientation histogram is generated. This process is illustrated in the following figure:

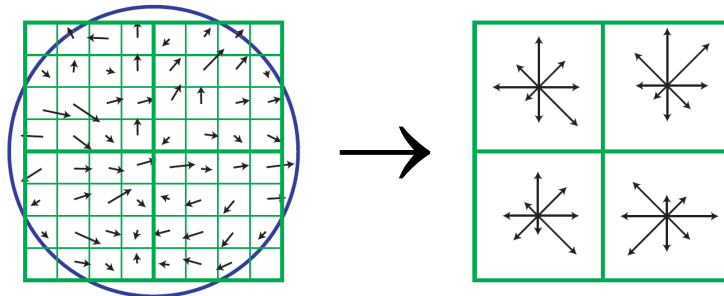


Figure 2.5: SIFT descriptor.

A Gaussian function (represented by the blue circle) with a standard deviation of one and a half the scale of the keypoint is used to weight the contribution of each sample. The purpose of this weighting function is to emphasize the gradients that are closer to the descriptor center, while reducing the contributions of the ones that are far from it.

These orientation histograms contain 8 bins each and summarize the local gradient information. The SIFT descriptor is computed from the values in these histograms, leading to a feature vector with $4 \times 4 \times 8 = 128$ elements.

Finally, the feature vector is changed to achieve invariance to changes in illumination. The influence of affine changes is reduced by normalizing the feature vector to unit length. Enhancement of invariance to non-linearities is achieved by decreasing the effects of large gradient magnitudes, thresholding the descriptor to 0.2 and then re-normalizing it to unit length. (The value of 0.2 was experimentally obtained by Lowe.)

2.5 Performance Evaluation

In this section, we describe some performance methods which are commonly used to evaluate the performance of classifier models and detection algorithms.

2.5.1 The Confusion Matrix

A confusion matrix provides a simple representation of the performance of a classifier model. It reports the number of *true positives*, *true negatives*, *false positives* and *false negatives*, comparing the true classifications, referred as *ground truth* or *gold standard*, with the predicted classifications of the classifier.

Table 2.4: The Confusion Matrix.

		Ground Truth	
		Positive	Negative
Test Outcome	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

The main diagonal of the confusion matrix contains the correct classifications, while the other fields represent model errors. Thus, for a perfect classifier, only the diagonal fields would be filled out.

A set of metrics may be derived from the confusion matrix and used to characterize the model's performance:

Accuracy (ACC)

The accuracy denotes the closeness of a measured value to the true value.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2.38)$$

Although this is one of the most used metrics, it is not fully reliable, since it does not consider the differences between the number of samples in the classes. For this reason, in the case of unbalanced data, this metric is not enough to represent the true performance of the model.

Sensitivity (SE)

Sensitivity, also called Recall or True Positive Rate (TPR), measures the proportion of actual positives that are correctly classified.

$$\frac{TP}{TP + FN} \quad (2.39)$$

Specificity (SP)

Specificity or True Negative Rate (TNR) measures the proportion of actual negatives that are correctly classified.

$$\frac{TN}{TN + FP} \quad (2.40)$$

False Positive Rate (FPR)

False Positive Rate ($1 - \textit{Specificity}$) measures the proportion of negative instances that are actually false negatives.

$$\frac{FP}{TN + FP} \quad (2.41)$$

Precision (PRC)

Precision, also called Positive Predicted Value, measures the proportion of positive instances that are actually true positives.

$$\frac{TP}{TP + FP} \quad (2.42)$$

2.5.2 The ROC Curve

A Receiver Operating Characteristic (ROC) curve is a two-dimensional graph that plots the TPR against the FPR for different decision thresholds of a parameter, providing a visually and intuitive analysis of the classifier model [24, 25].

Some regions of interest can be identified in a ROC graph, as shown in Fig. 2.6.

- The perfect classifier is denoted by the point on the top left corner, with maximum TPR (100%) and minimum FPR (0%).
- The random performance is represented by the diagonal line from the bottom left corner to the top right corner. Classifiers on this line produce the same number of TP and FP responses.
- In the conservative performance region (orange area), most instances are classified as negative. This means that there are few false positive errors, but the TPR is low.
- In the liberal performance (yellow area), models classify almost every observations as positive, producing few false negative errors and a high False Positive Rate.

- Below the random performance line (gray area), all the models have worse performances than the random performance.

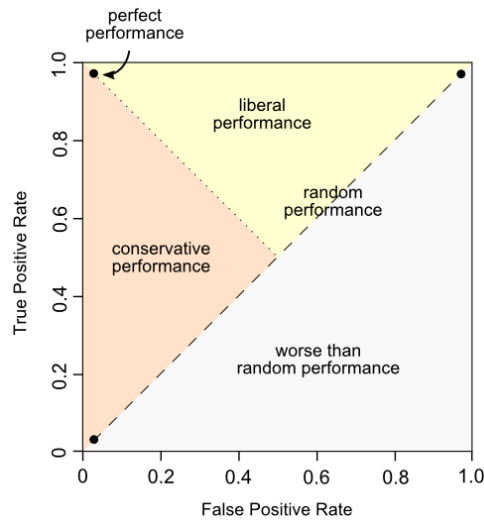


Figure 2.6: The ROC space.

The ROC curve shows a ranking of the classifiers based on their distances from the point denoting perfect performance. The closer the classifier is from the upper left corner, the higher the overall accuracy. Thus, the Area Under Curve (AUC) is a measure of the performance of the model.

The main advantage of the ROC curve over single measurements of TP/TP rates is that it avoids the loss of information caused by an arbitrary choice of the decision threshold. In fact, a ROC curve allows to compare the performance rates of a classifier for different decision thresholds, so the optimal operating point can be easily found with a quick visual analysis.

2.5.3 The PR Curve

Precision-Recall (PR) curves are an alternative picture of the algorithm's performance, plotting Precision against TPR. Just like in the ROC space, each point in the PR space represents one particular classifier, specified by a threshold value.

There is a strong difference between the visual representations of the curves in the ROC and PR spaces. The optimal operation point is represented by the upper-left-hand corner in the ROC space and by the upper-right-hand corner in the PR space. However, looking at both curves simultaneously allows to better evaluate the performance of a machine learning algorithm.

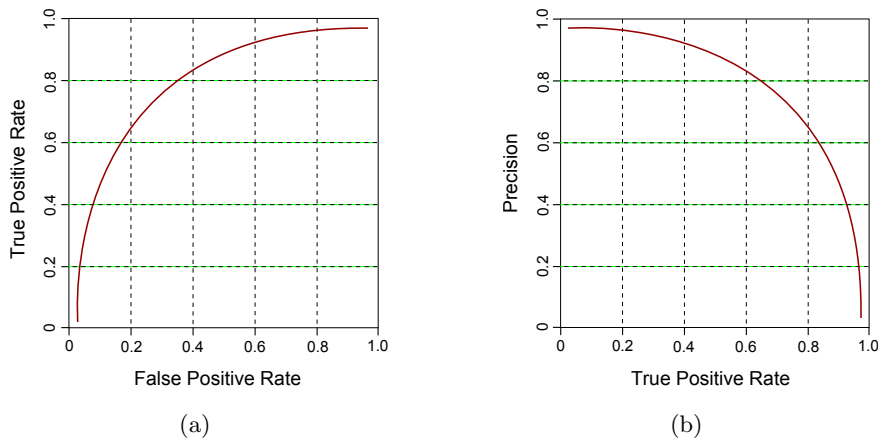


Figure 2.7: Visual comparison between a ROC curve (a) and a PR curve (b).

Although ROC and PR curves contain the same points, the PR space allows to observe differences between the models that are not visible in the ROC curve. This may be particularly important when dealing with highly skewed datasets since that, in these cases, ROC curves may lead to an optimistic view of the classification model's performance [25]. For example, if the number of negative samples largely exceeds the number of positive samples, SE and SP metrics, used in ROC analysis, may not properly reflect changes in the number of true and false positives. Since PRC does not consider the effect of the negative samples, these changes will be more realistically represented in the PR space.

PR curves are also very useful to evaluate the performance of a detection algorithm. ROC curves cannot be used in this framework, since it is not possible to measure TN instances and, therefore, compute SP values.

2.5.4 Cross-Validation

Cross-Validation is a statistical technique used in prediction tasks whose purpose is to assess the performance of a model in a dataset that is independent of the dataset used for training. Cross-validation concerns the estimation of the generalization power of an algorithm. Besides, it allows to compare the performance of two or more different models in order to find which one best fits for a certain dataset.

Here, we refer to some common cross-validation procedures.

Hold-Out Cross-Validation

This is the simplest Cross-Validation method. The data is randomly divided into two non-overlapped subsets: one is used as the training set to learn the classifier and the other as the testing set for the validation stage. The main disadvantage of hold-out technique is that the evaluation results highly depends on the choice of training and testing sets, which may lead to a high variance, specially in small datasets.

***k*-fold Cross-Validation**

The original dataset is partitioned into k folds and, then, the training and validation stages are repeated in k rounds. In each round, one of the k folds is held out to be used as testing set and the remaining $k - 1$ folds are put together to learn the model. The advantage of doing this is that every samples will be used both for training and validation, independently of how the data is divided. As the number of folds k increases, the variance of the evaluation results is reduced.

Leave-one-out Cross-Validation

This method is a special case of k -fold cross-validation, in which k equals to the number of samples in the original dataset. In each round, one single observation is retained for validation and all the others are used to train the model. Because it requires the training stage to be repeated many times, this method is computationally expensive.

Repeated random sub-sampling validation

The dataset is randomly split into a training and a testing sets and the process is repeated several times. The validation results are computed from averaging over the trials. When compared with k -fold cross-validation, the advantage of this method is that the split proportion of training and testing subsets is totally independent of the number of trials. However, it does not avoid overlapping of validation subsets, since some data samples may never be used in validation and other may be chosen more than once.

When analyzing the several listed cross-validation techniques, we may conclude that k -fold cross-validation is preferable to the hold-out method, since it reduces the variance related to the way the dataset is split. It is also preferable to the leave-one-out procedure because this one has a high computational cost.

Comparing k -fold cross-validation with repeated random subsampling, the second method eliminates the dependence between the number of trials and the proportion of training and testing subsets, but it does not avoid the overlap of subsets.

Since the SVM discriminator step of crypts detector provides information that will be used in the image classification stage, it is necessary to classify the whole dataset, not leaving any sample unclassified. Because of this, repeated random sub-sampling would not suit the problem, as it could leave some samples out of validation subsets. Therefore, we use k -fold cross validation.

Chapter 3

Image Segmentation

3.1 The segmentation task: The crypts

According to experts feedback, the presence of crypts in the tissue is one of the most relevant properties to take into account when making a diagnosis. The number of crypts, their shape, appearance and distribution over the tissue are determinant to distinguish between the several disease stages. Apart from this, crypts seem to be the most reliable top-level structures that may be easily observed in the tissue, so it becomes important to correctly identify and locate them. For these reasons, we decided to focus on image segmentation techniques in order to build a crypts detector.

Due to the great appearance variability of crypts, their segmentation is not a trivial task. As shown in Fig. 3.1, crypts are irregular structures with markedly different texture patterns, and contour boundaries that are often poorly defined. In fact, in many cases, it is difficult to accurately locate the exact frontier that separates the interior and exterior areas of the crypt, since there is a soft transition between these two regions.

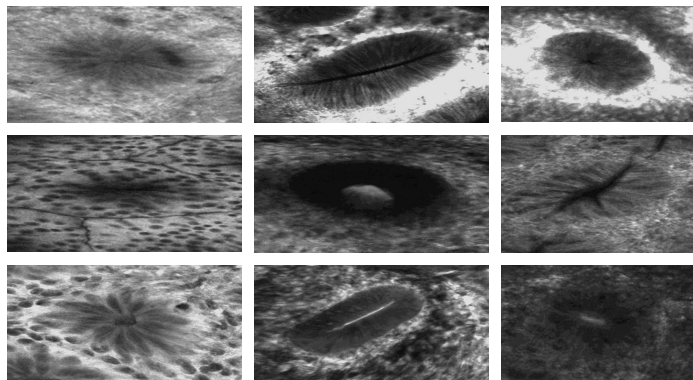


Figure 3.1: Crypts variability.

For the detection and segmentation of crypts, the proposed algorithm comprises three main steps, which are further explained in the following sections. We start by performing the detection of crypts' centers by searching for local maxima in a symmetry energy image (Section 3.2). Then, in the segmentation step, we use a canny filter to locate the contours of these structures and fit an ellipse to the edge points (Section 3.3). To deal with the problem of false detections, in the last stage, we use a binary SVM discriminator that will decide if each segmented region corresponds to a crypt (Section 3.4).

3.2 Centers Detection: A symmetry energy approach

The first stage of crypts detection is the location of interest points, which correspond to the centers of the crypts.

Despite crypts great variability, their shapes range from circular to elongate, reminding elliptical structures, which are symmetric objects. (Note that an ellipse has two lines of symmetry, which are its major and minor axis.) This fact supports the idea of using ellipse detection based on symmetry as an approach to identify crypts in the tissue.

3.2.1 Symmetry Energy - Background

An object which remains unchanged under a certain type of transformation is said to be symmetric. P. Kovesi [26] explored bilateral symmetry and asymmetry based on local intensity levels of an image signal. As explained in his work, symmetry is strictly related to the structure of objects. The fact that there is a certain degree of periodicity in the structure of a symmetric object led him to a frequency-based approach for the analysis of symmetry.

The extraction of local frequency information is done through multi-resolution wavelet analysis, by using a bank of filters tuned at different scales. Kovesi uses Log-Gabor wavelets, which are symmetric and anti-symmetric pairs of quadrature filters, to obtain amplitude and phase information from the image signal.

For an image signal I and a particular scale n , let M_n^e and M_n^o be, respectively, the even-symmetric (cosine) and the odd-symmetric (sine) wavelets.

Convolving the signal I with each pair of quadrature wavelets yields a response vector that may be written as:

$$[e_n(x), o_n(x)] = [I(x) * M_n^e, I(x) * M_n^o] . \quad (3.1)$$

Therefore, for each point x in the signal, there is an array of response vectors, one for each scale.

It is easily understandable that, at a point of symmetry, the output of the even-symmetric filter is larger than the output of the odd-symmetric filter. Accordingly, the opposite happens at a point of asymmetry. Thus, symmetry and asymmetry are quantified by the differences between the absolute values of these outputs, as follows:

$$Sym(x) = \frac{\sum_n [(|e_n(x)| - |o_n(x)|) - T]}{\sum_n A_n(x) + \epsilon} \quad (3.2)$$

(3.3)

$$ASym(x) = \frac{\sum_n [(|o_n(x)| - |e_n(x)|) - T]}{\sum_n A_n(x) + \epsilon} . \quad (3.4)$$

In these expressions, $A(x)$ denotes the amplitude of the transform, T is a noise compensation parameter and ϵ is a term that avoids division by zero.

3.2.2 Symmetry Energy - The Algorithm

Kovesi's Matlab code for symmetry detection is available online at [27]. The main input parameters of this implementation are the following:

- *nScale*: number of filter scales;
- *nOrient*: number of filter orientations;
- *minWaveLength*: wavelength of the smallest scale filter, which corresponds to the maximum center frequency, $f_{max} = \frac{1}{\lambda_{min}}$;
- *mult*: scaling between center frequencies of successive filters;
- *sigmaOnf*: parameter that controls filter bandwidth.

The algorithm returns:

- *phaseSym*: phase symmetry image, whose values vary between 0 and 1;
- *orientation*: orientation image, representing the orientations of maximum local symmetry energy;
- *totalEnergy*: un-normalized symmetry energy.

3.2.3 Symmetry Energy - Tuning parameters

The variation of the input parameters of Kovesi's algorithm strongly affects the response of the filter bank and, consequently, the detection results. Since it was impossible to guess which combination of input parameters would best suit to this particular case, it was necessary to make a tuning adjustment.

First, a set of images representing the crypts variability was chosen. Then, using this image set, the algorithm's parameters were varied and, for each combination of inputs, the recall and precision values in centers detection were computed. (Centers detection procedure and evaluation are further explained in sections 3.2.4 and 3.2.5, respectively.) From these results, the parameters which allowed to obtain higher recall and precision were selected.

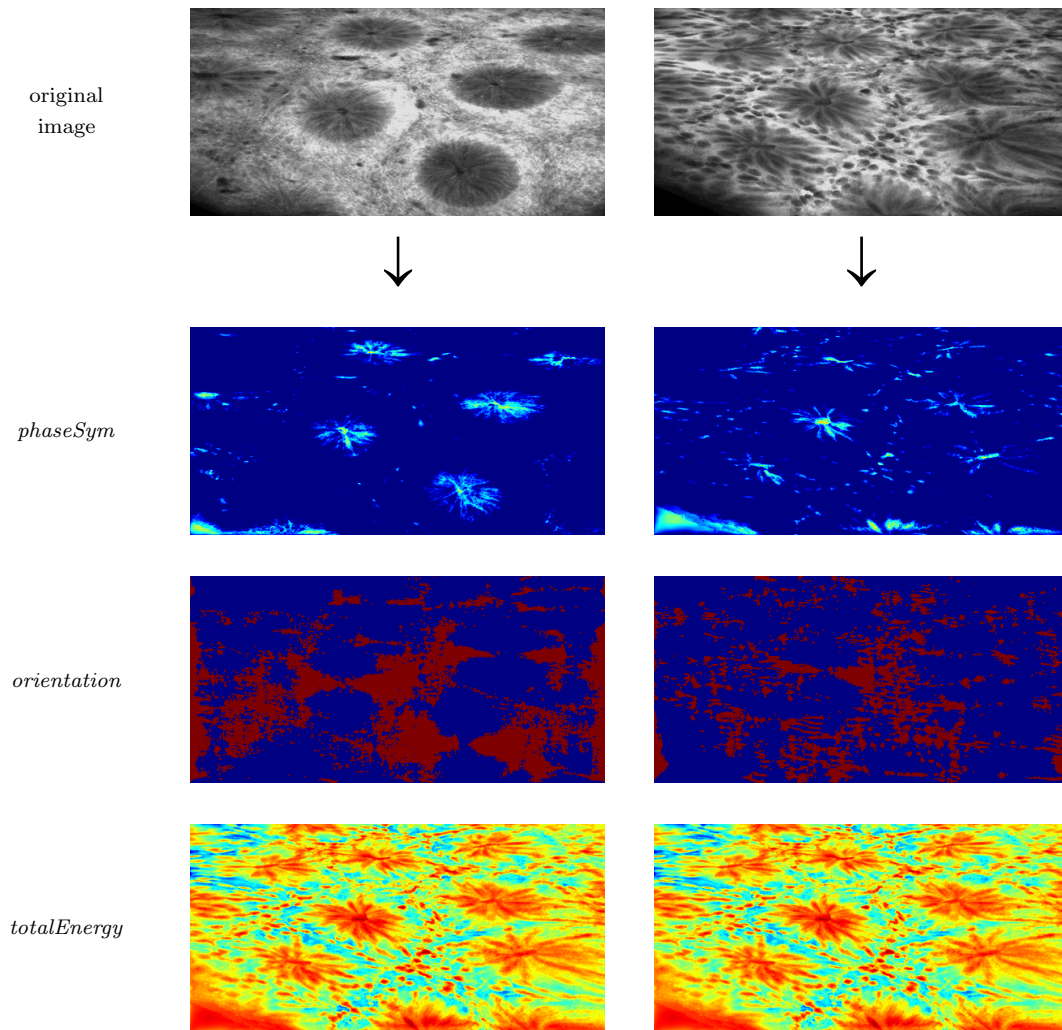


Figure 3.2: Results of Kovesi's algorithm: original CEM images and their corresponding outputs (*phaseSym*, *orientation* and *totalEnergy*), using the tuned parameters.

3.2.4 Local Maxima of Symmetry Energy

Based on the fact that crypts are structures with a high level of symmetry, the location of its centers is performed by computing local maxima of phase symmetry energy.

Local maxima are determined using a dilation morphological operation. Image dilation returns, for each pixel, the maximum value of a predefined neighborhood.

Note that, in this procedure, a smallest neighborhood produces a larger number of local maxima. Thus, this parameter may be adjusted to tune the detector, reducing or increasing the number of interest points.

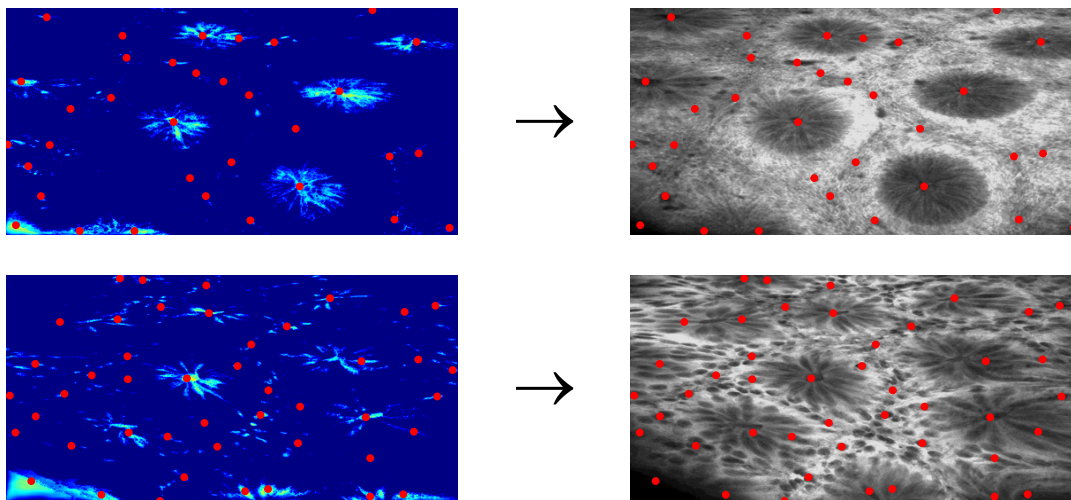


Figure 3.3: Result of crypts' centers detection. The red points correspond to the local maxima of symmetry energy.

3.2.5 Detection Results

In order to evaluate the success of centers detection, we needed to quantify the proportion of detected centers. Thus, it was required to define a criterion that clearly distinguishes a local maximum corresponding to a real center from a spurious point.

The chosen criterion is based on the distance from local maxima to the true centers of the crypts, as shown in the following figure:

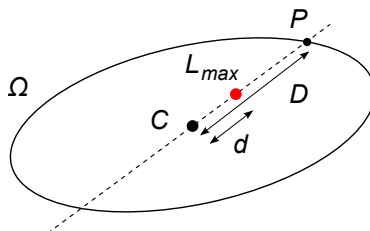


Figure 3.4: Criterion used to evaluate the success of centers detection stage.

Let Ω denote the ellipse that correctly segments the crypt. The real center C and the detected local maxima L_{max} define a line that intersects Ω in two points. Assume that P is the intersection point that is closer to L_{max} .

Let d and D be, respectively, the distances from C to L_{max} and from C to P . Given this, we define the ratio r as:

$$r = \frac{d}{D} . \quad (3.5)$$

If r is not too high, the local maximum is close enough to the real center of the crypt and may be accepted as a true detection. Based on this criterion, we may compute the detection results. As explained before in section 5.3.1, the number of determined local

maxima increases as the neighborhood of the dilation operator decreases. The following PR curve was obtained by varying the size of this neighborhood.

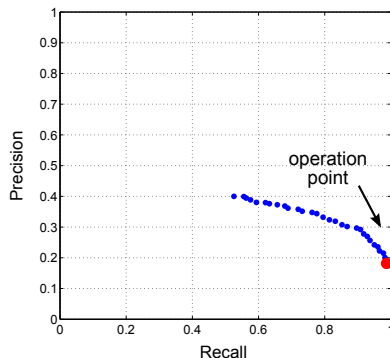


Figure 3.5: Centers detection: PR curve.

It is known that high recall with high precision is commonly difficult to achieve. As we can observe from the curve above, the keypoints detection procedure provides very low values of precision. This means that the detection stage not only catches the real centers of the crypts, but also several points that effectively denote symmetric structures in the image, but do not correspond to any crypt. Thus, in order to identify as many centers as possible, we have to deal with a lot of false detections.

Since there is no obvious solution to immediately discard the spurious points, we decided to follow a strategy of overdetection. The keypoints detector is tuned to catch every high symmetric location and the segmentation process is fully carried for all of the detected points. To face the problem of spurious keypoints, we then use a binary SVM discriminator that will decide if each one of the segmented regions corresponds to a crypt. This classification procedure will be explained further in section 3.4.

The following table shows the detection scores corresponding to the chosen operation point (marked in Fig. 3.5), which reflects the described overdetection approach. Note that, in classes 4 and 5, the scores are not computed, since there aren't regular crypts in these images.

Table 3.1: Detection Scores.

	Class 1	Class 2	Class 3	Class 4	Class 5	Global
Recall	0.985	0.989	0.991	-	-	0.987
Precision	0.344	0.301	0.081	-	-	0.183

3.3 Crypts Segmentation

3.3.1 Overview

Due to the elliptical shape of crypts, the most natural segmentation approach is to find an ellipse around each one of the detected keypoints. This is done by searching for boundary points around the center of the crypt and then fit a conic to them. The algorithm herein described was inspired and adapted from [28].

3.3.2 Basic Concepts on Conic Curves

In this section, we present a set of background concepts on conic curves which are useful to better understand the developed image segmentation algorithm.

The conic equation

Conic sections are curves obtained by intersecting a plane with a cone. A general conic is described by the following equation:

$$ax^2 + 2bxy + cy^2 - 2x - 2y + f = 0 . \quad (3.6)$$

This equation may be written in matrix notation as $x^T \Omega x = 0$, where x is a point lying on the conic in homogeneous coordinates and Ω is the real 3×3 symmetric matrix:

$$\Omega = \begin{bmatrix} a & b & d \\ b & c & e \\ d & e & f \end{bmatrix}_{3 \times 3} . \quad (3.7)$$

Typically, there are four different classes of non-degenerate conics, which can be distinguished by analyzing the discriminant, $b^2 - 4ac$.

- If $b^2 - 4ac < 0 \wedge a = 0 \wedge b = 0$, the conic equation represents a circle.
- If $b^2 - 4ac < 0$, the conic equation represents an ellipse.
- If $b^2 - 4ac = 0$, the conic equation represents a parabola.
- If $b^2 - 4ac > 0$, the conic equation represents a hyperbola.

Ellipse parameters

In this work, we are particularly interested in ellipses, due to the approximate elliptical shape of the crypts. An ellipse is fully described by its center, the length of major and minor axes and the rotation angle about the coordinate axes, as shown in the following figure:

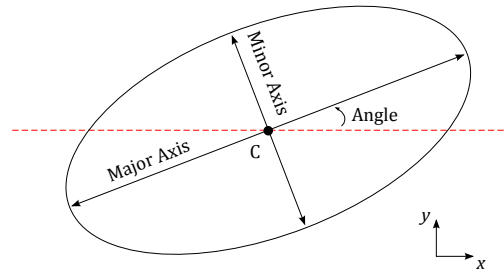


Figure 3.6: Ellipse parameters.

3.3.3 Brief Introduction to RANSAC

RANdom SAmples Consensus (RANSAC) [29] is a robust iterative procedure that allows to estimate the parameters of a mathematical model from a set of observations.

It assumes that the original dataset contains both inliers and outliers. Inliers are points whose distribution may be explained by the parameters of the model under evaluation. Outliers do not fit the model and may come from wrong observations or extreme values of noise.

This is a non-deterministic algorithm, since there is a probability related to the obtained result that depends on the number of inliers and outliers. This probability increases as more iterations are allowed during the procedure.

The RANSAC procedure may be described as follows:

- From the original dataset, randomly select n points needed for estimating the model.
- Estimate the fitting parameters from this subset.
- Check how many points from the total dataset fit the estimated model, i.e., are classified as inliers.
- If the number of inliers is big enough, accept the model.

These steps are repeated until a suitable result is produced or until the maximum number of iterations k is reached.

3.3.4 The Segmentation Algorithm

After locating the centers of the crypts, the next stage of the segmentation algorithm is to find the probable locations of the boundaries. This is done by applying an edge detector over the original image.

Edge Detection

Edge detection is one of the fundamental steps in image processing. It allows to identify quick and abrupt changes or discontinuities in image brightness, which characterize

objects' boundaries.

The directional changes in the intensity values in an image are measured by gradients, so most of the classical edge detectors are based on the analysis of image gradients. The idea is to construct a 2-D operator that is sensitive to large gradients. Convolving this filter with an image highlights edge information, returning zero values in gradient-uniform regions. Note that edge detection may turn out to be a tough task in case of noisy images, since both boundaries and noise are related to high frequencies.

There are several edge detection algorithms, some of them optimized to detect specific edge types, such as horizontal, vertical or diagonal edges. From the available techniques, Canny usually performs better, even in noisy scenarios, and it is often referred to as "the optimal edge detector" [30].

The Canny edge detector [31] uses a multi-stage algorithm, since its implementation involves a series of steps. Although we do not explore here the details of each one these steps, we pay a little attention to the first stage of the algorithm, which consists in noise reduction. Since the Canny operator is highly sensitive to noise, the use of a Gaussian blurring kernel strongly improves the final output. This is particularly important in our case because many images are noise-corrupted and have a lot of strong details that do not correspond to crypts' boundaries. For this reason, it is extremely important to correctly adjust the blurring degree, controlled by the standard deviation of the Gaussian filter, in order to smooth the image enough to eliminate noise but still preserving edge information. Note that the accuracy in the localization of the edges reduces as the size of the Gaussian is increased.

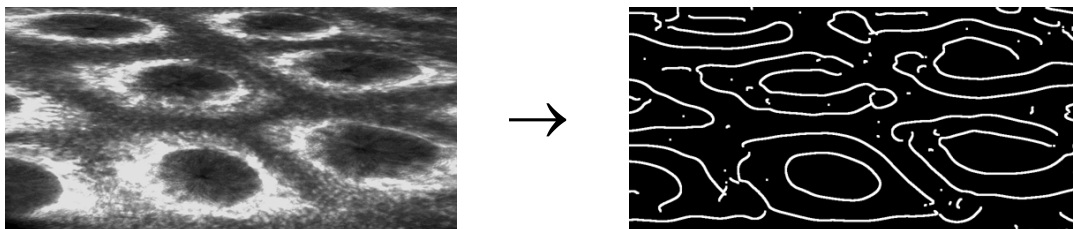


Figure 3.7: Result of the application of Canny filter over a CEM image.

Searching for boundary points

For each detected keypoint, let's consider a set of N radial equally spaced lines r_j , with $j = 1, 2, \dots, N$, that pass through it, corresponding to N radial directions. This radial neighborhood is scanned from the center to the periphery to find positive responses of the canny filter. However, instead of searching for edge points in the radial directions, we use a polar mapping, as proposed in [28], by changing from Cartesian coordinates $x = (x, y)$ to polar coordinates $\chi = (\rho, \theta)$.

We consider a searching circular patch in the original image whose center is the detected keypoint and whose radius is large enough to include the whole crypt. Let Ω be the curve

that surrounds this patch. First, a translation is applied to the patch to move its center from the keypoint coordinates (x_0, y_0) to the point $(0, 0)$, according to:

$$\Omega' = S^{-T} \Omega S^{-1}, \quad (3.8)$$

where the translation transformation S is given by:

$$S \sim \begin{pmatrix} 1 & 0 & -x_0 \\ 0 & 1 & -y_0 \\ 0 & 0 & 1 \end{pmatrix}. \quad (3.9)$$

The corresponding polar image is then generated using the inverse mapping, which maps the polar coordinates back into Cartesian coordinates:

$$\mathbf{x} \sim S^{-1} \begin{pmatrix} \rho \cos(\theta) \\ \rho \sin(\theta) \\ 1 \end{pmatrix}. \quad (3.10)$$

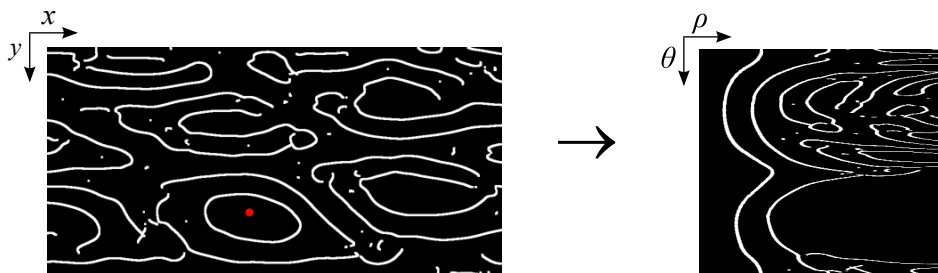


Figure 3.8: Generation of a polar image from the cartesian canny result.

In this polar image, the horizontal directions correspond to the original radial scanning lines r_j . Each horizontal line is then scanned from the left to the right until an edge point is reached. The detected edge points are mapped back from the polar space to the original Cartesian space using 3.10.

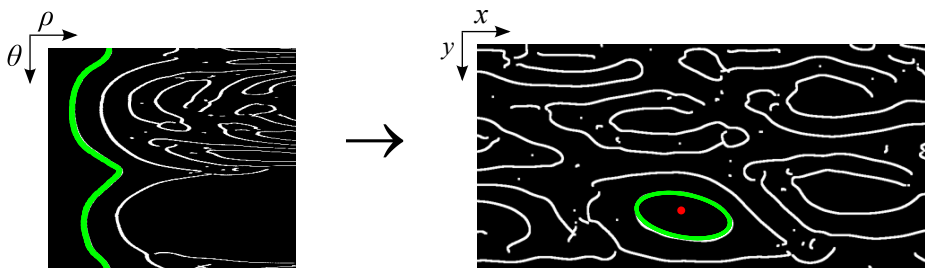


Figure 3.9: Inverse mapping from the polar to the cartesian space. The detected edge points are highlighted in green.

Ellipse Fitting and RANSAC

The equation of the ellipse that best fits the detected edge points is determined using a RANSAC procedure.

The choice of best estimation of the fitting ellipse is done based on the number of inliers of the model. It is important to note that, in this step, we added some constraints to the RANSAC procedure to obtain more realistic results. In each RANSAC iteration, an ellipse is discarded if:

- it does not contain the keypoint in its interior;
- its center is too distant from the original keypoint;
- it is too narrow. (Note that, although crypts' shapes are variable, the surrounding segmentation curves will never reach cases of extremely long and narrow ellipses.)

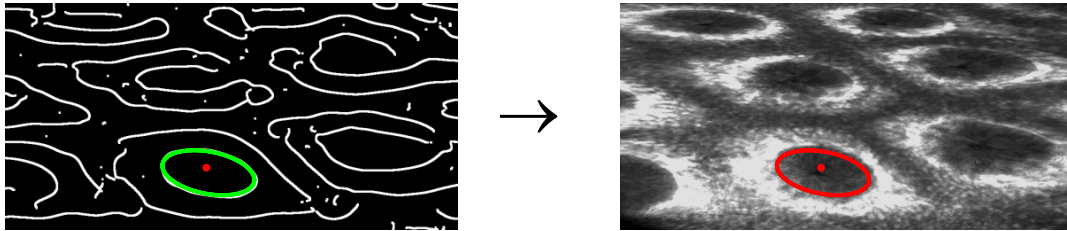


Figure 3.10: Fitting ellipse computed from the detected edge points, using a RANSAC procedure.

The edge detection algorithm returns a contour that tends to be within the real boundary of the crypt. The fitting ellipse is thus slightly enlarged to make sure that all the crypt area is enclosed by the curve.

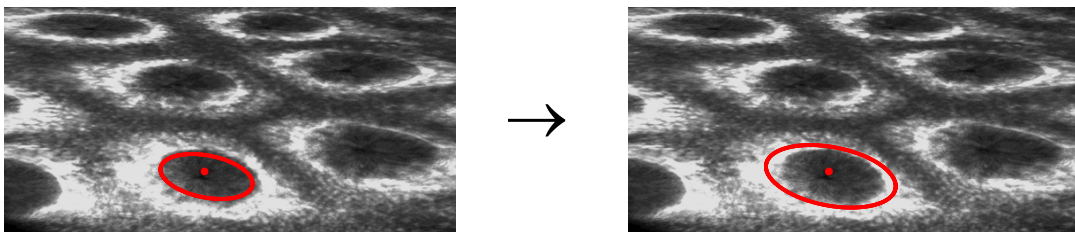


Figure 3.11: Final segmentation result.

3.3.5 Segmentation Results

At this point of the work, it was important to quantify the success of the segmentation procedure. Therefore, we had to establish a set of metrics to compare the segmentation curves provided by the crypts detector with the real database annotations. By doing this, it is possible to check if a certain ellipse may be accepted as a segmentation curve or not.

The criteria used in this comparative process are based in three different parameters:

- the distance between the centers of the two ellipses (Fig. 3.12(a));
- the overlap area between the two ellipses (Fig. 3.12(b));
- the difference between the rotation angles (Fig. 3.12(c)).

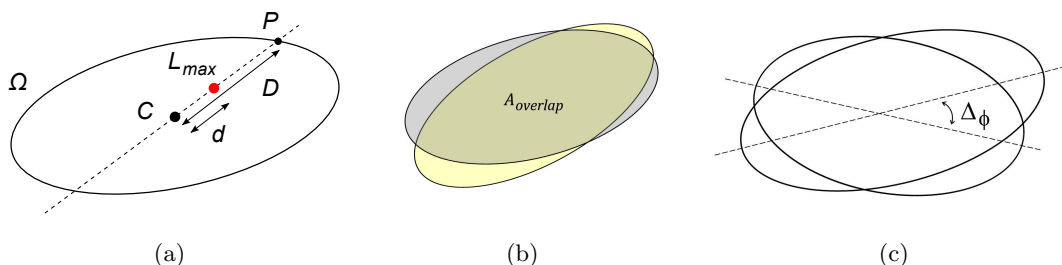


Figure 3.12: Criteria used to evaluate the performance of the segmentation algorithm.

The following table shows the partial segmentation scores, obtained in each one of the five classes, and the global scores, which refer to the whole dataset. Note that, in classes 4 and 5, the scores are not computed, since there aren't regular crypts in these images.

Table 3.2: Segmentation Scores.

	Class 1	Class 2	Class 3	Class 4	Class 5	Global
% segmented crypts	0.731	0.866	0.424	-	-	0.698

As we can see from the obtained results, the success of the segmentation algorithm is higher in classes 1 and 2 than in class 3. This is related to the tissue's appearance of the dataset images. In this two classes, there is a certain patterned arrangement of crypts. Besides, these structures show a more regular shape and their contours have stronger intensity transitions, which improves the canny's response.

3.4 SVM Discriminator

In this stage, we built a binary SVM classifier to discard the spurious segmented crypts. This will handle the problem of false detections and incorrectly segmented crypts and, therefore, improve the precision of crypts detector.

We focus on the design of local description schemes that encode the appearance of the segmented regions. The goal is to create a feature vector that is discriminative enough to distinguish image regions that correspond to a crypt from those in which these structures are not present.

3.4.1 The Normalization Step

Before feature extraction, we perform a normalization of the segmented regions.

We achieve scale and affine invariance in local features detection and description to provide a more robust classification. Scale invariance eliminates the differences in the size of crypts, related to the appearance variability of these structures. Affine invariance not only allows to reach shape normalization, but also reduces the slant effect that occurs if the endoscope is not positioned perpendicularly to the tissue surface during image acquisition.

Since crypts are circular symmetric structures, rotation invariance is not necessary.

Affine and Scale Invariance

Affine invariance considers non-uniform scaling, which means that there is invariance even if the scales in the orthogonal directions are different.

The anisotropic shape of a local structure can be computed from the second moment matrix μ , also known as the auto-correlation matrix [32, 33]. The second moment matrix provides a measure of the local isotropy, given by the ratio between its eigenvalues:

$$Q = \frac{\lambda_{min}}{\lambda_{max}}, \quad (3.11)$$

where λ_{min} and λ_{max} represent, respectively, the minimum and maximum eigenvalues of μ . Q ranges from 0 to 1, with 1 for a perfect isotropic structure.

The goal is to transform the anisotropic region into an isotropic region by using the normalization matrix A that maps the computed ellipse into a circumference. The estimation of A is based on the properties of the second moment matrix. Thus, A is given by:

$$A = \sqrt{\mu}. \quad (3.12)$$

Warping the image patch according to this affine transformation results in a normalized isotropic patch in which the computed features will be affine invariant.

Scale invariance is reached by mapping the several segmented elliptical regions to equally sized circumferences.

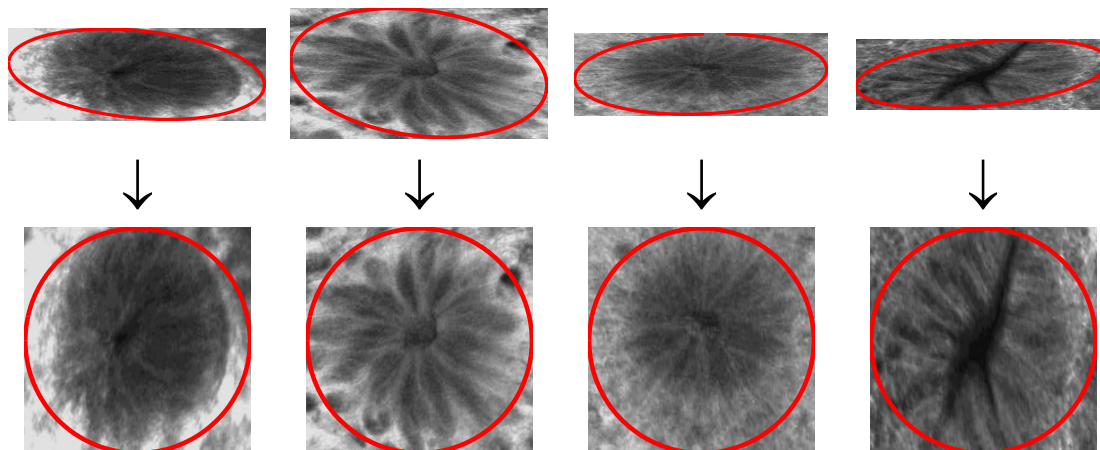


Figure 3.13: The normalization step: Affine and Scale Invariance.

3.4.2 Selection of Features

For each one of the detected keypoints, texture features are computed over the normalized patch. We used texture features derived from image histograms, GLCM, GLRLM and Laws Energy. We also compute the popular SIFT descriptor [7], that usually achieves high performances in several object recognition and matching tasks.

In addition, we experiment with a new descriptor, specifically designed for this problem, that we dubbed Radial Gradient Descriptor.

The Radial Gradient Descriptor

Crypts do not have well-defined boundaries. In fact, these structures do not always show strong edges and, sometimes, they can only be identified by small intensity or texture variations between their interior and exterior regions. We created a radial descriptor based on image gradients in order to catch these variations and identify a transition region that may correspond to the boundary of the crypts and, thus, encode the shape of these structures.

To assure that crypts' contour is included in the patch used for feature extraction, it is not enough to consider only their interior area. For this reason, given the ellipse that defines a keypoint region, we slightly enlarge it before the normalization step, as shown in Fig. 3.14 (a) and (e).

Then, we compute a radial profile of gradient magnitudes, by summing magnitude values over 360 degrees around the center. For robustness against image contrast and brightness changes [7], this radial profile is normalized by the square root of sum of squares of all the original values, according to the L2-Norm.

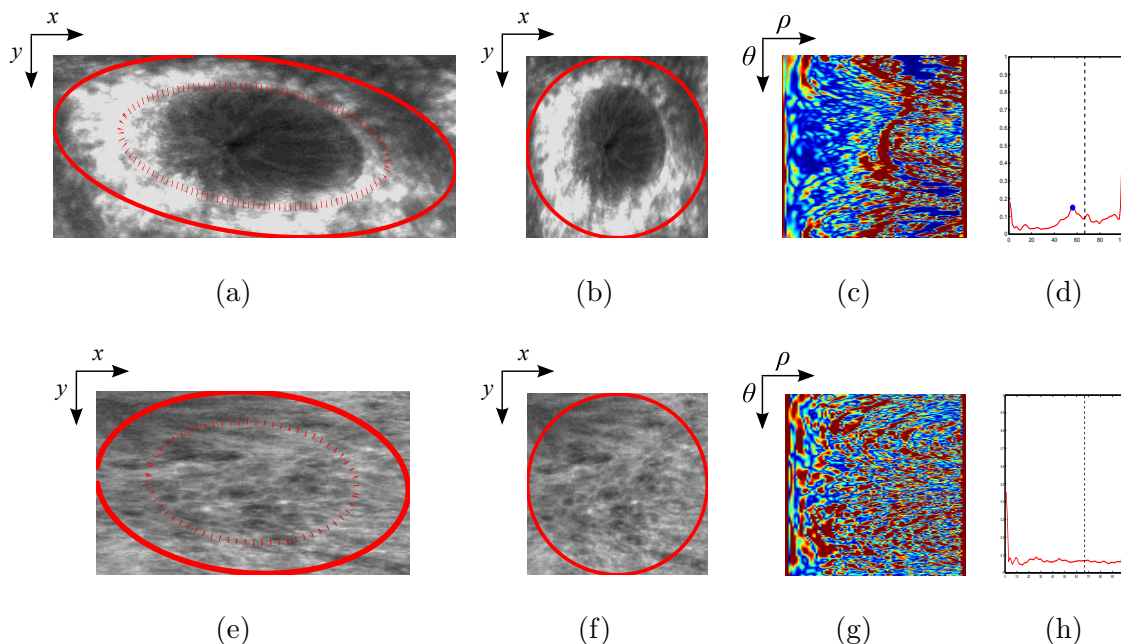


Figure 3.14: The figures show the radial gradient profiles of two different segmented regions: one corresponds to a crypt (first row) and the other one to a false detection (second row). The ellipses computed in the segmentation stage, denoted by the red dashed ellipses in (a) and (e), are enlarged to include a possible boundary. The patches are then normalized to achieve affine and scale invariance, as shown in (b) and (f), and the radial gradient magnitudes are computed from a polar mapping (see (c) and (g)). Finally, the radial gradient profiles, (d) and (h), are obtained by averaging the radial gradient magnitude values across rows of the radial images. Note that, in (d) and (h), the black dashed vertical lines correspond to the red dashed ellipses in (a) and (e).

If we look carefully at Fig. 3.14 d), we can observe that the radial profile does not behave the same in the interior and the exterior regions of the crypt (corresponding to the left and right sides of the black dashed line, respectively). Since crypts are usually darker than other tissue's areas, the magnitude values of the descriptor are higher in the outside of the crypt. It is even possible to identify a slight peak of magnitude, marked by the blue dot, which corresponds to the boundary of the crypt. This does not happen in Fig. 3.14 (h), in which the differences between the interior and exterior regions of the computed ellipse are much more subtle.

3.4.3 SVM Discriminator Scores

Although desirable, achieving both high recall and precision is not easy. Since we are using an overdetection approach, we pretend to find a combination of features that allows to discard as many false detections as possible. In this overdetection context, a high discriminative classifier with a reasonable recall is preferable to a high recall but low precision model. We compare the performances of textures features (Histogram Moments, GLCM, GLRLM and Law’s Energy) and SIFT descriptor with the proposed radial gradient profile to find which one best fits this particular classification problem.

We use the RFB kernel and 10-fold cross-validation to classify the whole dataset. The SVM parameters σ (spread of RFB kernel) and C (regularization term) of the models were previously adjusted using a two-layer grid-search, as proposed in [34].

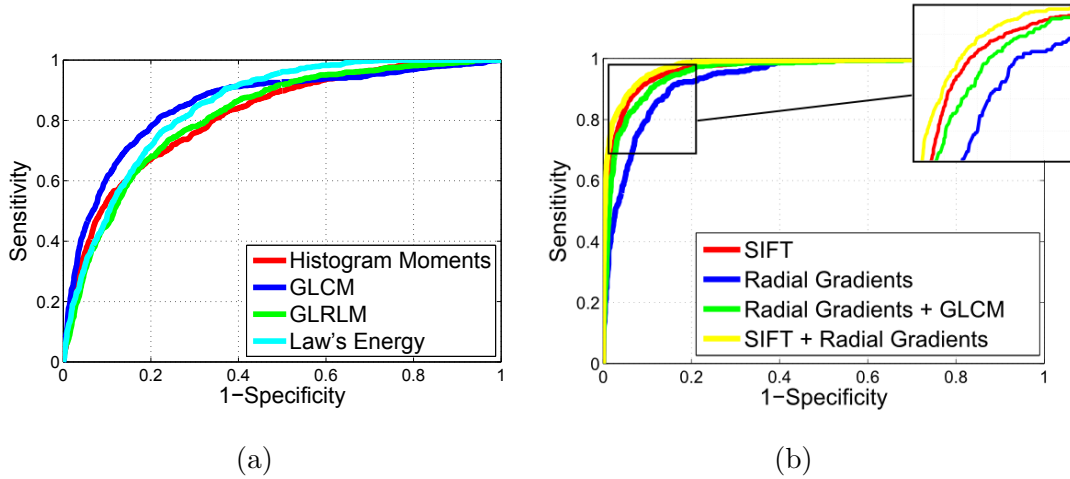


Figure 3.15: SVM Discriminator Performance: The graphs show the 10-fold cross-validation averaged ROC curves of the evaluated models. In (a), we compare the several texture features; in (b) we compare the remaining descriptors.

Table 3.3: Performance metrics of the several models: The table shows the mean values of ACC, SE, SP, PRC and AUC (computed by averaging the results over the 10-fold cross-validation rounds).

	ACC	SE	SP	PRC	AUC
Histogram Moments	0.745	0.716	0.749	0.296	0.818
GLCM	0.831	0.688	0.852	0.406	0.848
GLRLM	0.750	0.750	0.750	0.305	0.820
Law’s Energy	0.762	0.778	0.760	0.322	0.841
SIFT	0.911	0.904	0.912	0.616	0.967
Radial Gradients	0.841	0.888	0.834	0.440	0.935
Radial Gradients + GLCM	0.905	0.883	0.908	0.584	0.959
Radial Gradients + SIFT	0.950	0.751	0.981	0.862	0.977

All the evaluated texture features show poor performance levels. The precision values are very low, which means that there is still a great number of false positives and, thus, the classifiers based on these features are not able to discard the spurious detections. Keeping in mind that we pretend to find a high discriminative model, the results show that GLCM performs better than the other evaluated texture features, since it has higher precision (although the recall is slightly decreased). However, at least when used alone, none of these features is enough discriminative to deal with this particular classification problem.

Our radial gradient profile clearly outperforms texture features both in terms of SE and SP, but the precision is still under the clinical desired level. Fig. 3.16 shows some examples of TP, TN, FP and FN. Looking carefully at these examples allows to conclude about the limitations of the proposed descriptor. From TP and FP examples, we can observe that it is possible to confuse a crypt descriptor to a false detection's descriptor. In fact, there several symmetric structures in the tissue that do not correspond to a crypt but that behave similarly in what concerns the radial gradient profile. The TN and FN examples show that the differences in gradient magnitudes inside and outside a crypt are not always sufficient to identify it, since the gradients transition may be too subtle.

However, when combined with texture information (GLCM), the obtained performance values of the produced model are very similar to SIFT's, with the advantage that our descriptors are much simpler and easier to compute.

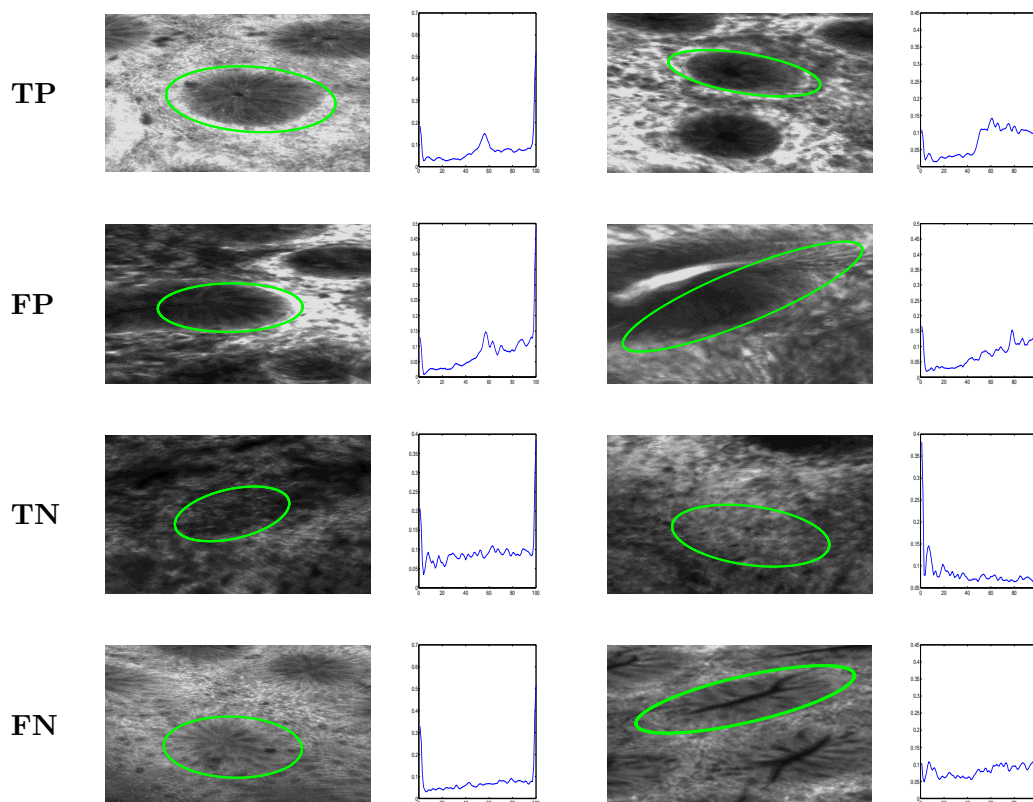


Figure 3.16: Examples of TP, FP, TN and FN obtained by the classifier trained with our proposed radial gradient profile.

As we can see from Table 3.3, when combining the proposed radial gradient profile with the SIFT descriptor, the robustness of the classifier increases significantly. Although the recall is slightly decreased (0.751), the precision is much higher (0.862) than when using any one of these two descriptors alone. For this reason, we consider that this combination of features is the most suitable to use in this task. The next set of figures show the results before and after the SVM discriminator.

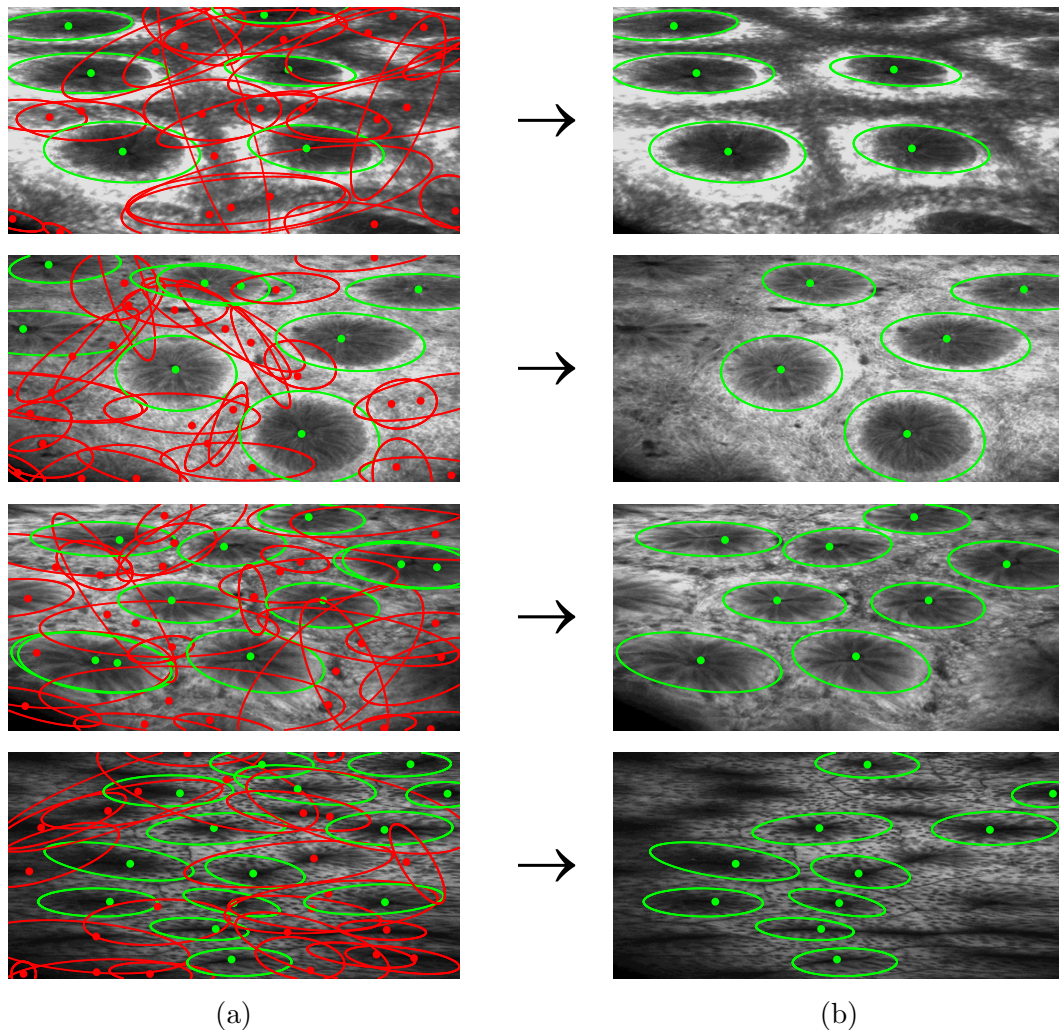


Figure 3.17: Results obtained before (a) and after (b) the SVM discriminator when using the radial gradient profile and SIFT as features. In (a), the green ellipses correspond to correctly segmented crypts and the red ellipses denote spurious detections and incorrectly segmented crypts. As we can see from these examples, the classifier is able to discard the spurious segmented crypts, due to its good precision. However, since the recall is not as high as the desired, it also discards some true positives. (In fact, some of the green ellipses from (a) are eliminated by the SVM and do not appear in (b).)

3.5 Global Evaluation of Crypts Detector

The intermediate results of the three stages of crypts detector have were already been presented and discussed in the present chapter. In this section, those results are summarized in order to provide a global evaluation of the entire proposed algorithm.

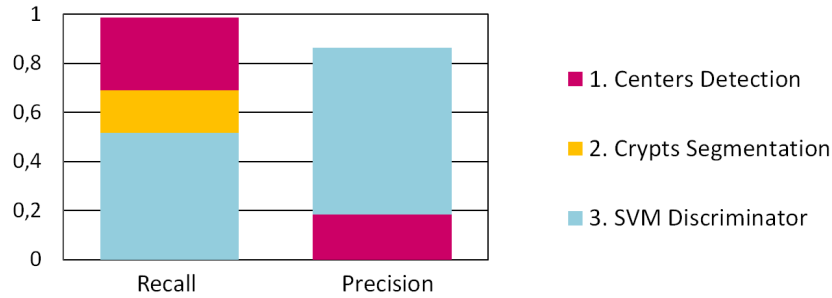


Figure 3.18: Global evaluation of crypts detector performance. The bars show the cumulative results of the three stages of the proposed algorithm: centers detection, crypts segmentation and SVM discriminator.

In the first step of crypts detector, the obtained recall is very close to 1, which means that almost every crypts are detected. However, the precision is extremely low and, thus, there is a great number of false positives. These results reflect the chosen overdetection approach.

In the segmentation stage, the results clearly denote that the algorithm is not robust enough to deal with the huge crypts variability. However, given the difficulty of the problem, the current success rates are quite acceptable.

The third step of the algorithm definitely needs to be improved, specially in what concerns the TPR of the classifier. The chosen combination of features provides high precision values, but not the desired recall rates. However, as already explained before, this situation is preferable to a high recall but low precision classifier; otherwise the final output of the algorithm would return both crypts and spurious segmented regions. From a clinical point of view, this would not be very useful to the doctors, since the main purpose is to highlight relevant structures in the tissue.

Chapter 4

Image Classification

4.1 Classification Strategy

Our five-class problem is a complex multi-class classification task, due to the intra and inter class variability and to the lack of a clearly defined taxonomy for CEM images, as explained previously in Section 1.4.

The available dataset is not yet large enough to get a representative sample of the five different stages of illness. In fact, in classes 4 and 5, the few number of collected images does not provide sufficient examples of the visual intra-class variability to build a robust multi-class classifier. Because of this, instead of facing the whole multi-class problem at once, we stick to a binary classification task, which, from a medical diagnosis point-of-view, consists in distinguishing two main stages of IBD: *low* and *high probability of pathology*.

4.2 Dataset binary subdivision

According to experts, the tissue-level organization is one of the most relevant properties that allow to distinguish between a healthy and a pathological stage. In healthy stages, there is a certain orderly arrangement of tissue's architecture and it is possible to recognize several histological structures. Pathological stages are frequently characterized by a strong disorganization of the tissue's appearance, which is sometimes close to a chaotic situation.

Based on this, we propose to split the data into two main subsets: subset 1 (classes 1/2), in which there is a certain patterned arrangement of crypts, and subset 2 (classes 3/4/5), when the tissue's appearance is closer to a disordered stage. This data division is accomplished by a standard SVM classifier that decides to which subset each input image belongs.

4.2.1 Features used

As pointed out by medical experts, the spatial organization of crypts in CEM images is highly discriminative to distinguish between disease stages. In fact, when looking at some examples of subsets 1 and 2 1.2, it is possible to observe that there are markedly differences in the tissue's organization.

To include information about tissue's organization, the SVM is applied on a feature vector that encodes the number of crypts in each image, provided by the crypts detector, and their arrangement in the tissue (*lattice*), computed by using Delaunay Triangulation. Given a set of P points in a plane, a Delaunay triangulation is such that each triangle contains any point in its interior. The *lattice* is then described by the mean distance between the several crypts and the corresponding standard deviation.

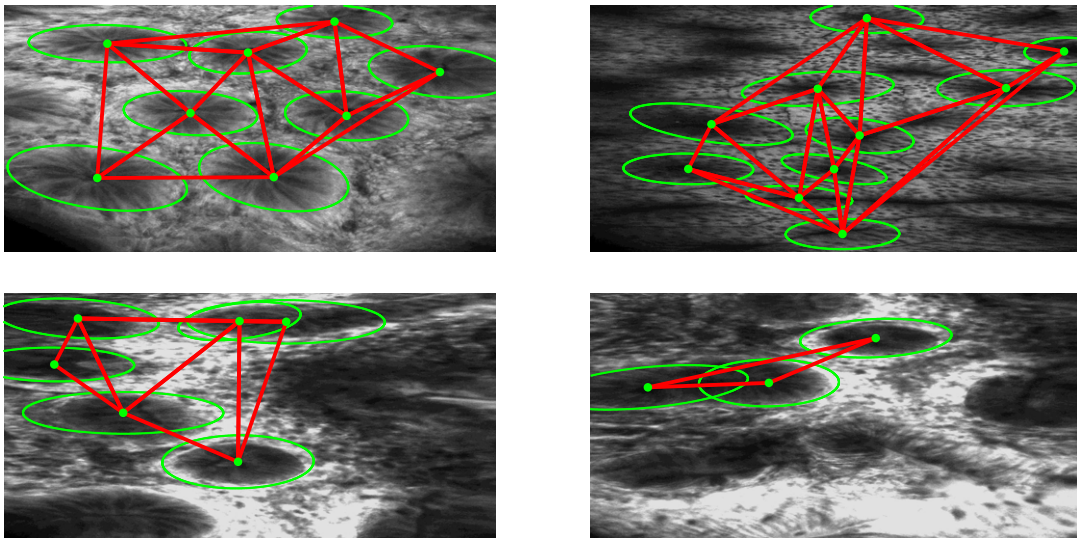


Figure 4.1: Features used for image classification. In each image, it is possible to observe the number of detected crypts and the *lattice*, computed by Delaunay Triangulation.

4.2.2 Image Classification Scores

We compare the features we proposed based on experts interpretation of CEM images with SIFT descriptor [5–7]. As proposed in [5, 6], we quantize SIFT features into visual words and encode them in image histograms. Then, instead of performing the retrieval, we use these histograms to train the SVM classifier. The results are shown in Fig. 4.2 and Table 4.2.

Like in the SVM discriminator stage, we use the RFB kernel and 10-fold cross-validation to classify the whole dataset. The SVM parameters σ (spread of RFB kernel) and C (regularization term) were adjusted using a two-layer grid-search, as proposed in [34].

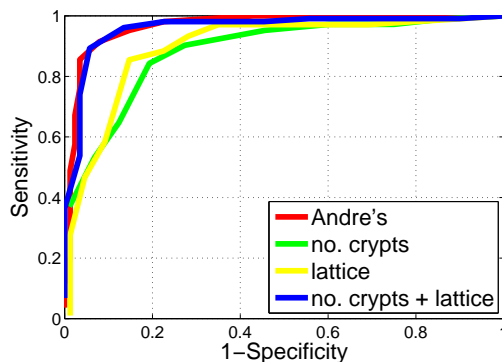


Figure 4.2: Image Classifier Performance: The graphs show the 10-fold cross-validation averaged ROC curves of the evaluated features.

Table 4.1: Performance metrics of the several features: The table shows the mean values of ACC, SE, SP, PRC and AUC (computed by averaging the results over the 10-fold cross-validation rounds).

	ACC	SE	SP	PRC	AUC
Andre's	0.932	0.972	0.883	0.919	0.968
no. crypts	0.860	0.865	0.854	0.877	0.910
<i>lattice</i>	0.853	0.826	0.886	0.898	0.914
no. crypts + <i>lattice</i>	0.932	0.933	0.932	0.946	0.960

The obtained results show that the proposed histological-based features contain reliable information to use in this classification task. When used alone, the number of crypts and *lattice* do not achieve extremely good performances. However, combining both features leads to a classifier that performs very close to Andre's method in our available dataset. We consider this result quite satisfactory, since we are using conceptually simpler features, which are much easier to compute and require much less storage space.

Chapter 5

Conclusions and Future Work

This work presents preliminary research toward building a CEM image classification system. Unlike other relevant CEM image classification studies [5, 6] that blindly apply techniques widely used in several application areas, we build our system based on the physicians interpretation of the histological architecture of the GI tract.

Given the complexity of the classification problem and the small size of the available dataset, we consider the current results encouraging toward further research. The segmentation of relevant structures in the tissue is not only important to the classification stage, but also provides useful imaging analysis information, which is helpful for clinical evaluation and for the diagnosis. The image classifier's performance shows that the histological appearance of CEM images provide reliable information to build a discriminative and robust classification system.

We consider that some technical improvements still need to be carried out to improve the current results. In what concerns the image segmentation stage, we believe that there is possible to refine the proposed algorithm. The ellipse fitting process based on RANSAC turns out to be relatively slow due to the poor accuracy of the detection of edge points around the centers. Thus, future research directions should focus on the improvement of the algorithm both in terms of computational time and accuracy.

Another future step is to enhance the discriminative power of the features used in the SVM discriminator. The purpose is to develop specific crypts descriptors to achieve both high recall and precision rates.

The ultimate goal will concern the development of an automatic classification system to distinguish the several IBD stages. This will involve the enlargement of the available dataset to obtain numerous examples of the several classes. Besides, the enhancement of CEM images taxonomy will be crucial to identify relevant tissue's properties that may be used to train new image classifiers.

Bibliography

- [1] G. D. De Palma, “Confocal laser endomicroscopy in the ”in vivo” histological diagnosis of the gastrointestinal tract.” *World J Gastroenterol*, vol. 15, no. 46, pp. 5770–5, 2009. [Online]. Available: <http://www.biomedsearch.com/nih/Confocal-laser-endomicroscopy-in-vivo/19998496.html>
- [2] A. Hoffman, M. Goetz, M. Vieth, P. R. Galle, M. F. Neurath, and R. Kiesslich, “Confocal laser endomicroscopy: technical status and current indications.” *Endoscopy*, vol. 38, no. 12, pp. 1275–83, 2006. [Online]. Available: <http://www.biomedsearch.com/nih/Confocal-laser-endomicroscopy-technical-status/17163333.html>
- [3] D. Gheonea, A. Saftoiu, T. Ciurea, C. Popescu, C. Georgescu, and A. Malos, “Confocal laser endomicroscopy of the colon.” *J Gastrointestin Liver Dis*, vol. 19, no. 2, pp. 207–11, 2010.
- [4] A. Brito, “Protocol and support infrastructure for the creation of an annotated database of images by confocal endomicroscopy.” Faculty of Sciences and Technology, University of Coimbra, Master’s thesis, 2011.
- [5] B. André, T. Vercauteren, A. M. Buchner, M. B. Wallace, and N. Ayache, “A smart atlas for endomicroscopy using automated video retrieval,” *Medical Image Analysis*, vol. 15, no. 4, pp. 460–476, 2011.
- [6] B. André, T. Vercauteren, A. Perchant, A. M. Buchner, M. B. Wallace, and N. Ayache, “Introducing space and time in local feature-based endomicroscopic image retrieval,” in *Proceedings of the First MICCAI international conference on Medical Content-Based Retrieval for Clinical Decision Support*, ser. MCBR-CDS’09. Berlin, Heidelberg: Springer-Verlag, 2010, pp. 18–30.
- [7] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 91–110, Nov. 2004. [Online]. Available: <http://dx.doi.org/10.1023/B:VISI.0000029664.99615.94>
- [8] A. Gidudu, G. Hulley, and T. Marwala, “Image classification using svms: One-against-one vs one-against-all,” *CoRR*, vol. abs/0711.2914, 2007.

- [9] S. Marsland, *Machine Learning: An Algorithmic Perspective*, 1st ed. Chapman & Hall/CRC, 2009.
- [10] E. Alpaydin, *Introduction to Machine Learning*, 2nd ed. The MIT Press, 2010.
- [11] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Min. Knowl. Discov.*, vol. 2, no. 2, pp. 121–167, Jun. 1998. [Online]. Available: <http://dx.doi.org/10.1023/A:1009715923555>
- [12] R. Sahak, W. Mansor, L. Y. Khuan, A. Yassin, A. Zabidi, and F. Rahman, "Choice for a support vector machine kernel function for recognizing asphyxia from infant cries," in *Industrial Electronics Applications, 2009. ISIEA 2009. IEEE Symposium on*, vol. 2, oct. 2009, pp. 675–678.
- [13] M. Aly, "Survey on multi-class classification methods," 2005.
- [14] J. Milgram, M. Cheriet, and R. Sabourin, "'One Against One' or 'One Against All': Which One is Better for Handwriting Recognition with SVMs?" in *Tenth International Workshop on Frontiers in Handwriting Recognition*, G. Lorette, Ed., Université de Rennes 1. La Baule (France): Suvisoft, Oct. 2006, <http://www.suvisoft.com> Université de Rennes 1. [Online]. Available: <http://hal.inria.fr/inria-00103955>
- [15] C.-W. Hsu and C.-J. Lin, "A comparison of methods for multiclass support vector machines," *IEEE Transactions on Neural Networks*, vol. 13, no. 2, pp. 415–425, 2002.
- [16] A. Liaw and M. Wiener, "Classification and regression by randomforest," *R News*, vol. 2, no. 3, pp. 18–22, 2002. [Online]. Available: <http://CRAN.R-project.org/doc/Rnews/>
- [17] L. B. Statistics and L. Breiman, "Random forests," in *Machine Learning*, 2001, pp. 5–32.
- [18] A. Bosch, A. Zisserman, and X. Munoz, "Image classification using random forests and ferns," in *Proceedings of the 11th International Conference on Computer Vision, Rio de Janeiro, Brazil, 2007*.
- [19] L. Breiman and L. Breiman, "Bagging predictors," in *Machine Learning*, 1996, pp. 123–140.
- [20] F. Moosmann, E. Nowak, and F. Jurie, "Randomized clustering forests for image classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 9, pp. 1632–1646, Sep. 2008. [Online]. Available: <http://dx.doi.org/10.1109/TPAMI.2007.70822>
- [21] M. Partio, B. Cramariuc, M. Gabbouj, and A. Visa, "Rock texture retrieval using gray level co-occurrence matrix."

- [22] X. Tang, "Texture information in run-length matrices." *IEEE Transactions on Image Processing*, vol. 7, no. 11, pp. 1602–1609, 1998. [Online]. Available: <http://www.ncbi.nlm.nih.gov/pubmed/18276225>
- [23] M. Sharma and S. Singh, "Evaluation of texture methods for image analysis," in *Intelligent Information Systems Conference, The Seventh Australian and New Zealand 2001*, nov. 2001, pp. 117 – 121.
- [24] L. Hamel, "Model assessment with roc curves," *The Encyclopedia of Data Warehousing and Mining*, vol. The Encycl, 2001. [Online]. Available: <http://homepage.cs.uri.edu/faculty/hamel/pubs/hamel-roc.pdf>
- [25] J. Davis and M. Goadrich, "The relationship between precision-recall and roc curves," in *Proceedings of the 23rd international conference on Machine learning*, ser. ICML '06. New York, NY, USA: ACM, 2006, pp. 233–240. [Online]. Available: <http://doi.acm.org/10.1145/1143844.1143874>
- [26] P. Kovesi, "Symmetry and asymmetry from local phase," in *Tenth Australian Joint Conference on Artificial Intelligence*, 1997, pp. 2–4.
- [27] P. D. Kovesi, "MATLAB and Octave functions for computer vision and image processing," Centre for Exploration Targeting, School of Earth and Environment, The University of Western Australia, available from: <http://www.csse.uwa.edu.au/~pk/research/matlabfns/>.
- [28] R. Melo, J. P. Barreto, and G. Falcao, "A new solution for camera calibration and real-time image distortion correction in medical endoscopy-initial technical evaluation," *IEEE Trans. Biomed. Engineering*, vol. 59, no. 3, pp. 634–644, 2012.
- [29] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, Jun. 1981. [Online]. Available: <http://doi.acm.org/10.1145/358669.358692>
- [30] R. Maini, "Study and comparison of various image edge detection techniques," *Image Processing*, vol. 147002, no. 3, pp. 1–12, 2009. [Online]. Available: <http://www.cscjournals.org/csc/manuscriptinfo.php?ManuscriptCode=72.73.72.79.44.48.52.99>
- [31] J. Canny, "A computational approach to edge detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 8, no. 6, pp. 679–698, Jun. 1986.
- [32] K. Mikolajczyk and C. Schmid, "Scale and affine invariant interest point detectors," *International Journal of Computer Vision*, vol. 60, no. 1, pp. 63–86, 2004. [Online]. Available: <http://lear.inrialpes.fr/pubs/2004/MS04>
- [33] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. V. Gool, "A comparison of affine region detectors," *International*

- Journal of Computer Vision*, vol. 65, no. 1/2, pp. 43–72, 2005. [Online]. Available: <http://lear.inrialpes.fr/pubs/2005/MTSZMSKG05>
- [34] C.-W. Hsu, C.-C. Chang, and C.-J. Lin, “A practical guide to support vector classification,” Department of Computer Science, National Taiwan University, Tech. Rep., 2003. [Online]. Available: <http://www.csie.ntu.edu.tw/~cjlin/papers.html>
- [35] C. Goncalves and V. Bairos, *Histologia. Texto e Imagens*. Imprensa da Universidade de Coimbra, 2010.
- [36] B. André, T. Vercauteren, A. M. Buchner, M. B. Wallace, and N. Ayache, “A smart atlas for endomicroscopy using automated video retrieval,” *Medical Image Analysis*, vol. 15, no. 4, pp. 460–476, 2011.
- [37] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, 2nd ed. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 2001.
- [38] T. K. Ho, “The random subspace method for constructing decision forests,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 8, pp. 832–844, Aug. 1998. [Online]. Available: <http://dx.doi.org/10.1109/34.709601>
- [39] M. Antunes, “Stereo from symmetry: The signal processing approach,” Faculdade de Ciencias e Tecnologia da Universidade de Coimbra, Tech. Rep., 2009.
- [40] D. G. Lowe, “Object recognition from local scale-invariant features,” in *Proceedings of the International Conference on Computer Vision-Volume 2 - Volume 2*, ser. ICCV ’99. Washington, DC, USA: IEEE Computer Society, 1999, pp. 1150–. [Online]. Available: <http://dl.acm.org/citation.cfm?id=850924.851523>