



Facultade de Informática
Departamento de Computación

PHD THESIS

Advancing the diagnosis of dry eye syndrome:
development of automated assessments of
tear film lipid layer patterns

Beatriz Remeseiro López
March 2014

PhD advisors:
Manuel F. González Penedo
Antonio Mosquera González

March 19, 2014
UNIVERSIDADE DA CORUÑA

FACULDADE DE INFORMÁTICA
Campus de Elviña s/n
15071, A Coruña (Spain)

Copyright notice:

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording and/or otherwise without the prior permission of the authors.

To my parents and sisters

Acknowledgments

German philosopher Georg Wilhelm Friedrich Hegel said “*Nothing great in the World has been accomplished without passion*”. This passion is what characterizes all my oldest sister does. She has opened all the doors for me since we were children, and so she has also opened me the door to research. Since I am in the middle (which is the way madness lies!), I also have a youngest sister. She is a little bit grumpy, but I love her since her smile makes life easier. And what to say about my parents, who have supported me since I started my career some years ago. Their faith in me and their love have made it possible that my dream is about to come true.

I would like to thank my PhD advisors for their hundreds of useful suggestions to the title of this thesis. I needed more than patience to please both of you, and of course myself! All jokes aside, thanks to Manuel F. González Penedo for introducing me to the research world and letting me belong to your research group. Thank you for being a breath of fresh air and giving me the opportunity to carry out this thesis, after I had come to an impasse with my previous topic. Antonio Mosquera González, thank you for your careful guidance, your smart advices and your creative drawings. I have resigned myself to never having such a beautiful notebook full of perfect improvised doodles.

And now is the turn to write about the VARPA group. What does VARPA mean? It means research and interminable meetings, jokes and fun, and almost everything you can imagine to make these years simply unforgettable. Thank you for sharing with me these long and sometimes hard years. A special thanks goes to Noelia for being my “almost third advisor”, and solving my weird problems.

This thesis would be not possible without the people from the Faculty of Optics and Optometry of the University of Santiago de Compostela (Spain), who had the tempting idea of automatizing their manual processes. Thank you for bringing me closer to the clinical world and providing me with hundreds of images. In addition, thank you to all the people from the School of Sciences of the University of Minho (Portugal), who have also collaborated in this line of research by means of productive talks and imaging tasks.

From the Iberian Peninsula to Scotland, my acknowledgments travel there to express my gratitude to Professor Alan Tomlinson. Thank you for allowing me to enjoy a research stay in Glasgow Caledonian University (UK), and work with one of the most recognized specialists in dry eye research. Thank you to Dr Katherine Oliver and Eilidh Martin for their warm welcome, and for the wonderful experience of working together. I would like to highlight the altruistic collaboration of Katherine's husband with his key contribution to the title of this thesis. And of course, I want to mention the great memories that I have of the three months I spent there, and all the people that I had the opportunity to meet. *"Guys, just don't die!"*.

Last but far from least is a heartfelt thank you to all my friends for all the great moments that we have enjoyed together in these last few years. Specially, the ones who have walked with me the path to becoming a doctor.

*"And now, the end is near
And so I face the final curtain
...
And did it my way!"*
Frank Sinatra

Abstract

Dry eye syndrome is a symptomatic disease which affects a wide range of population, and has a negative impact on their daily activities. Its diagnosis is a difficult task due to its multifactorial etiology, and so there exist several clinical tests. One of these tests is the evaluation of the interference patterns of the tear film lipid layer. Guillon designed an instrument known as Tearscope Plus which allows clinicians to rapidly assess the lipid layer thickness, and also defined a grading scale composed of five categories. The classification into these five patterns is a difficult clinical task, especially with thinner lipid layers which lack color and/or morphological features. Furthermore, the subjective interpretation of the experts via visual inspection may affect the classification, and so a high degree of inter- and also intra- observer variability can be produced. The development of a systematic, objective computerized method for analysis and classification is thus highly desirable, allowing for homogeneous diagnosis and relieving the experts from this tedious task.

The proposal of this research is the design of an automatic system to assess the tear film lipid layer patterns through the interpretation of the images acquired with the Tearscope Plus. On the one hand, a global methodology is presented to assess the tear film lipid layer by automatically classifying these images into the Guillon categories. The process is carried out using texture and color models, and machine learning algorithms. Then, this global methodology is optimized through the reduction of its computational complexity. Dimensionality reduction techniques are used in order to diminish the memory/time requirements with no degradation in performance. On the other hand, a local methodology is also presented to create tear film maps, which represent the local distribution of the lipid layer patterns over the tear film. The different automated assessments proposed save time for experts, and provide unbiased results which are not affected by subjective factors.

Resumen

El síndrome de ojo seco es una enfermedad sintomática que afecta a un amplio rango de la población, y tiene un impacto negativo en sus actividades diarias. Su diagnóstico es una tarea difícil debido a su etiología multifactorial, y por eso existen varias pruebas clínicas. Una de esas pruebas es la evaluación de los patrones interferenciales de la capa lipídica de la película lagrimal. Guillon diseñó un instrumento denominado Tearscope Plus para evaluar el grosor de la película lagrimal de forma rápida, y también definió una escala de clasificación compuesta de cinco categorías. La clasificación en uno de esos cinco patrones es una tarea clínica difícil, especialmente con las capas lipídicas más finas que carecen de características de color y/o morfológicas. Además, la interpretación subjetiva de los expertos mediante una revisión visual puede afectar a la clasificación, pudiendo producirse un alto grado de inter- e intra- variabilidad entre observadores. El desarrollo de un método sistemático y objetivo para análisis y clasificación es altamente deseable, permitiendo un diagnóstico homogéneo y liberando a los expertos de esta tediosa tarea.

La propuesta de esta investigación es el diseño de un sistema automático para evaluar los patrones de la capa lipídica de la película lagrimal mediante la interpretación de las imágenes obtenidas con el Tearscope Plus. Por una parte, se presenta una metodología global para evaluar la capa lipídica de la película lagrimal mediante la clasificación automática de estas imágenes en una de las categorías de Guillon. El proceso se lleva a cabo mediante el uso de modelos de textura y color, y algoritmos de aprendizaje máquina. A continuación, esta metodología global se optimiza mediante la reducción de su complejidad computacional. Se utilizan técnicas de reducción de la dimensión para disminuir los requisitos de memoria/tiempo sin una degradación en su rendimiento. Por otra parte, se presenta una metodología local para crear mapas de la película lagrimal, que representan la distribución local de los patrones de la capa lipídica sobre la película lagrimal. Las diferentes evaluaciones automáticas que se proponen ahorran tiempo a los expertos, y proporcionan resultados imparciales que no están afectados por factores subjetivos.

Resumo

O síndrome de ollo seco é unha enfermidade sintomática que afecta a un amplo rango da poboación, e ten un impacto negativo nas súas actividades diarias. O seu diagnóstico é unha tarefa difícil debido á súa etioloxía multifactorial, e por iso existen varias probas clínicas. Unha desas probas é a avaliación dos patróns interferenciais da capa lipídica da película lagrimal. Guillon deseñou un instrumento denominado Tearscope Plus para avaliar o grosor da película lagrimal de forma rápida, e tamén definiu unha escala de clasificación composta de cinco categorías. A clasificación nun deses cinco patróns é unha tarefa clínica difícil, especialmente coas capas lipídicas máis finas que carecen de características de cor e/ou morfolóxicas. Ademais, a interpretación subxectiva dos expertos mediante una revisión visual pode afectar á clasificación, podendo producirse un alto grao de inter- e intra- variabilidade entre observadores. O desenvolvemento dun método sistemático e obxectivo para análise e clasificación é altamente desexable, permitindo un diagnóstico homoxéneo e liberando aos expertos desta tediosa tarefa.

A proposta desta investigación é o deseño dun sistema automático para avaliar os patróns da capa lipídica da película lagrimal mediante a interpretación das imaxes obtidas co Tearscope Plus. Por unha parte, preséntase unha metodoloxía global para avaliar a capa lipídica da película lagrimal mediante a clasificación automática destas imaxes nunha das categorías de Guillon. O proceso é levado a cabo mediante o uso de modelos de textura e cor, e algoritmos de aprendizaxe máquina. A continuación, esta metodoloxía global é optimizada mediante a redución da súa complexidade computacional. Utilízanse técnicas de redución da dimensión para diminuír os requisitos de memoria/tempo sen unha degradación no seu rendemento. Por outra parte, preséntase unha metodoloxía local para crear mapas da película lagrimal, que representan a distribución local dos patróns da capa lipídica sobre a película lagrimal. As diferentes avaliacións automáticas que se proponen aforran tempo aos expertos, e proporcionan resultados imparciais que non están afectados por factores subxectivos.

Contents

1	Introduction	1
1.1	Tear film	2
1.1.1	Lipid layer	3
1.1.2	Aqueous layer	4
1.1.3	Mucous layer	4
1.2	Dry eye syndrome	4
1.2.1	The classification of dry eye syndrome	4
1.2.2	The epidemiology of dry eye syndrome	6
1.3	Clinical tests for dry eye diagnosis	8
1.3.1	Quantitative tear film tests	8
1.3.2	Qualitative tear film tests	8
1.4	Lipid layer pattern assessment	10
1.4.1	Specular reflection	11
1.4.2	Interference phenomena	11
1.4.3	Tearscope Plus	12
1.5	Image datasets	16
1.5.1	Image acquisition	16
1.5.2	Illumination conditions	17
1.5.3	VOPTICAL_I1 dataset	17
1.5.4	VOPTICAL_Is dataset	18
1.5.5	VOPTICAL_R dataset	19
1.6	Thesis	19
2	Tear film assessment	21
2.1	Research methodology	22
2.2	Location of the region of interest	22
2.2.1	Experimental study	23
2.3	Feature vector	24

2.3.1	Color analysis	24
2.3.2	Texture analysis	26
2.3.3	Definition of the feature vector	31
2.3.4	Experimental study	31
2.4	Classification	37
2.4.1	Machine learning algorithms	37
2.4.2	Experimental study	41
2.5	Conclusions	43
3	Dimensionality reduction	47
3.1	Feature extraction	47
3.1.1	Principal component analysis	48
3.1.2	Experimental study	48
3.2	Feature selection	50
3.2.1	Filters	50
3.2.2	Experimental study	52
3.3	Cost-based feature selection	59
3.3.1	mC-ReliefF	60
3.3.2	Experimental study	61
3.4	Conclusions	65
4	Tear film distribution maps	67
4.1	Optimal window size	68
4.1.1	Experimental study	68
4.2	Research methodology	69
4.2.1	Location of the region of interest	71
4.2.2	Feature vector	71
4.2.3	Soft classification	73
4.2.4	Definition of the tear film map	73
4.2.5	Post-processing	78
4.2.6	Experimental study	80
4.3	Conclusions	93
5	Conclusions	97
5.1	Further research	99
A	Experimental results	101
A.1	Texture analysis	101
A.2	Classification	109

A.3	Principal component analysis	118
B	Co-occurrence features	121
B.1	Statistical measures	122
C	Estimating the accuracy of classifiers	125
C.1	k -fold cross-validation	125
C.2	Leave-one-out cross-validation	126
D	Comparing classifiers: statistical analysis	127
D.1	The Lilliefors test for normality	128
D.2	The ANOVA test	129
D.3	The Tukey's method for multiple comparison	129
E	Evaluation of tear film lipid layer classification	131
E.1	Methodology	132
E.2	A case of study	134
E.2.1	Data acquisition: tear film images	134
E.2.2	Class binarization techniques	135
E.2.3	Feature selection: filters	136
E.2.4	Classification: machine learning algorithms	137
E.2.5	Performance measures	137
E.2.6	Decision-making: multiple-criteria decision-making methods	138
E.2.7	Conflict handling: Spearman's rank correlation coefficient	139
E.3	Experimental results	140
E.3.1	Results	140
E.3.2	Conflict handling results	142
E.4	Conclusions	144
F	Publications and other mentions	145
G	Resumen	151
G.1	Aspectos clínicos	152
G.2	Tesis	153
G.3	Conclusiones	154
	Bibliography	157

List of Figures

1.1	Structure of the eye	1
1.2	Structure of the tear film	3
1.3	Major etiological causes of dry eye	6
1.4	Schirmer and phenol read thread tests	9
1.5	Appearance of tear meniscus height	9
1.6	Appearance of the superficial lipid layer	10
1.7	Tear break-up time	10
1.8	Not invasive tear break-up time	11
1.9	Hemispherical light source	12
1.10	Optical diagram: interference phenomena	13
1.11	Tearscope Plus	14
1.12	Patterns with different areas	17
1.13	Images with different illuminations	18
2.1	Methodology to assess tear film lipid layer patterns	22
2.2	Location of the ROI	23
2.3	Examples of ROIs	24
2.4	Simplified texture patterns	27
2.5	Feature using grayscale images	31
2.6	Feature vector using the Lab color space	32
2.7	Feature vector using opponent colors	32
2.8	Experimental procedure: color and texture methods	33
2.9	Experimental procedure: classifiers	41
3.1	Experimental procedure: PCA	49
3.2	Experimental procedure: feature selection filters	52
3.3	Experimental procedure: mC-ReliefF	62
3.4	Error/cost plots and Pareto front	64

4.1	Heterogeneity of the tear film lipid layer	67
4.2	Local windows and their feature vectors	69
4.3	Experimental procedure: optimum window size	69
4.4	Window size vs accuracy	70
4.5	Methodology for tear film distribution maps	70
4.6	Location of ROI	72
4.7	Post-processing step	80
4.8	Experimental procedure 1	81
4.9	Tear film maps: decision voting system	82
4.10	Experimental procedure 2	83
4.11	Tear film maps: weighted voting system and seeded region growing	84
4.12	Comparison between optometrists	85
4.13	Qualitative analysis: weighted voting system ω_1, ω_2	87
4.14	Qualitative analysis: weighted voting system th	89
4.15	Qualitative results: seeded region growing	91
D.1	Methodology for the statistical analysis of classifiers	128
E.1	Methodology of evaluation	132
E.2	Methodology applied to tear film lipid layer classification	134

List of Tables

1.1	Color and thickness of the lipid layer patterns	13
1.2	Appearance and thickness of the lipid layer patterns	15
2.1	Best combinations for color and texture analysis	36
2.2	The most competitive classifiers	44
3.1	Parameter configurations for texture analysis	53
3.2	Number of selected features	54
3.3	Mean test classification accuracy	55
3.4	Robustness	55
3.5	Feature computing time	56
3.6	Features within distances and components	57
3.7	Set of the 27 features using CFS	58
3.8	Performance measures for co-occurrence features analysis	59
3.9	Error, time and number of features	65
4.1	Performance measures: weighted voting system ω_1, ω_2	88
4.2	Performance measures: weighted voting system th	90
4.3	Performance measures: seeded region growing	92
4.4	Average time to create tear film maps	93
A.1	Butterworth filters: SVM classification accuracy	102
A.2	Gabor filters: SVM classification accuracy	103
A.3	The discrete wavelet transform: SVM classification accuracy	104
A.4	Markov random fields: SVM classification accuracy	104
A.5	Co-occurrence features: SVM classification accuracy	106
A.6	Two method combinations: SVM classification accuracy	107
A.7	Three method combinations: SVM classification accuracy	108
A.8	Four method combinations: SVM classification accuracy	109
A.9	Five method combination: SVM classification accuracy	109

A.10 Butterworth filters: classification accuracy	110
A.11 Butterworth filters: ANOVA results	111
A.12 Gabor filters: classification accuracy	112
A.13 Gabor filters: ANOVA results	112
A.14 The discrete wavelet transform: classification accuracy	113
A.15 The discrete wavelet transform: ANOVA results	114
A.16 Markov random fields: classification accuracy	115
A.17 Markov random fields: ANOVA results	115
A.18 Co-occurrence features: classification accuracy	116
A.19 Co-occurrence features: ANOVA results	117
A.20 PCA using different variances and grayscale	118
A.21 PCA using different variances and opponent colors	119
A.22 PCA using different variances and Lab	120
E.1 Features selected	141
E.2 TOP 10 alternatives ranked by TOPSIS	142
E.3 TOP 10 alternatives ranked by GRA	142
E.4 TOP 10 alternatives ranked by VIKOR	143
E.5 MCDM rankings and values	143
E.6 Weights and normalized weights	143
E.7 Weighted MCDM rankings and values	144

List of Algorithms

3.1	Pseudo-code: ReliefF	61
4.1	Pseudo-code: decision voting system	74
4.2	Pseudo-code: weighted voting system	76
4.3	Pseudo-code: seed search	78
4.4	Pseudo-code: region growing	79

Chapter 1

Introduction

The eyes are undoubtedly some of the most delicate, sensitive and complex organs we possess (Miller, 1969). They present us with the window through which we view the world, and are responsible for four fifths of all the information our brain receives. For this reason, we probably rely on our eyesight more than any other sense. The surface of the eye, known as the ocular surface, consists of the cornea and the conjunctiva (see Figure 1.1). It is an extraordinary and vital component of vision. As a mucosa, it is protected by the immune system that uses innate and adaptive effector mechanisms present in the tear film.

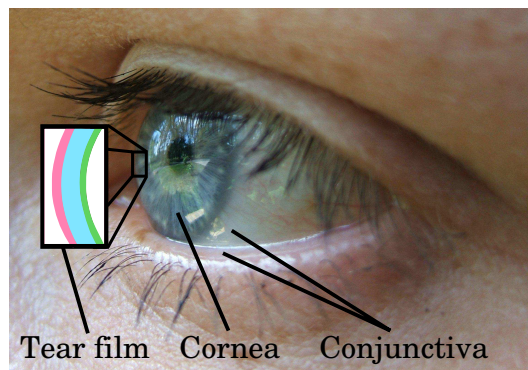


Figure 1.1: Structure of the eye.

Tears are secreted from the lachrymal gland and distributed by blinking to form the tear film of the ocular surface (Pflugfelder et al., 1998). The tear film is responsible for wetting the ocular surface, which is the first line of defense, and is also essential for clear visual imaging (Rieger, 1992). Its outer layer, known as *tear film lipid layer*, is composed of a polar phase with surfactant properties overlaid by a

nonpolar phase. It is the thinnest layer of the tear film and is mainly secreted by the meibomian glands, embedded in the upper and lower tarsal plates (K. K. Nichols, Nichols, & Mitchell, 2004).

A quantitative or qualitative change in the normal lipid layer has a negative effect on the quality of vision measured as contrast sensitivity, and on the evaporation of tears from the ocular surface (Rolando, Iester, Macrí, & Calabria, 1998). Actually, it has been shown that a substantial tear evaporation caused by alterations of the lipid layer is characteristic of the *evaporative dry eye* (EDE). This disease leads to irritation of the ocular surface, and is associated with symptoms of discomfort and dryness. It is a common complaint among middle-aged and older adults, and affects a wide range of population (Lemp et al., 2007b): between 10% and 20% of the population, although in Asian populations this percentage may be raised up to 33%. It affects specially among contact lens users, and worsens with age. The current work conditions, such as computer use, have increased the proportion of people with EDE (Lemp et al., 2007a).

1.1 Tear film

The tear film covers the exposed anterior surface of the eye and is essential for the execution of its functions, such as the maintenance of a healthy and functional visual system. Its main important functions are (Korb, 2002):

Optical function. The tear film fills in the irregularities of the corneal epithelium, and so provides a perfect, smooth, regular optical surface. So, an absence of the tear film provokes blur vision.

Lubrication function. It allows to minimize the friction between eyelid margins and palpebral conjunctiva during blinking.

Cleaning function. The tear film, together with blinking action, removes debris and desquamated epithelial cells from the epithelium.

Antimicrobial function. The tear film is the first line of defense against ocular surface infection. It contains proteins, such as lysozyme or lactoferrin, which inhibit microbiological contamination.

Nutritional function. Corneal surface must be avascular to guarantee its transparency, so the nutrition is driven by the tear film. Oxygen from the ambient air dissolves in the tear fluid and is transferred to the corneal epithelium.

The total volume of the tear film is $7.0 \pm 2.0 \mu\text{l}$ with a thickness ranging from $6 - 10 \mu\text{m}$. Along the upper and lower lids, it forms a tear meniscus or marginal tear strips. This represents 70% of the total volume of tear fluid within the palpebral aperture (Larke, 1997). A small proportion lies beneath the eyelids between the palpebral and bulbar conjunctiva, and the remainder covers the cornea and the exposed bulbar conjunctiva (Korb, 2002).

The tear film is a matrix-like structure composed of water, electrolytes, immunoglobulins, antimicrobial molecules and mucins. Wolff provided the classical description of the precocular tear film as a three-layered structure (Wolff, 1954), which consists of an anterior lipid layer, an aqueous layer, and a deep mucin layer (see Figure 1.2). Each of these layer plays a different role towards the formation and stability of the structure. In this process, not only the quality and quantity of each layer are important, but also their relationship.

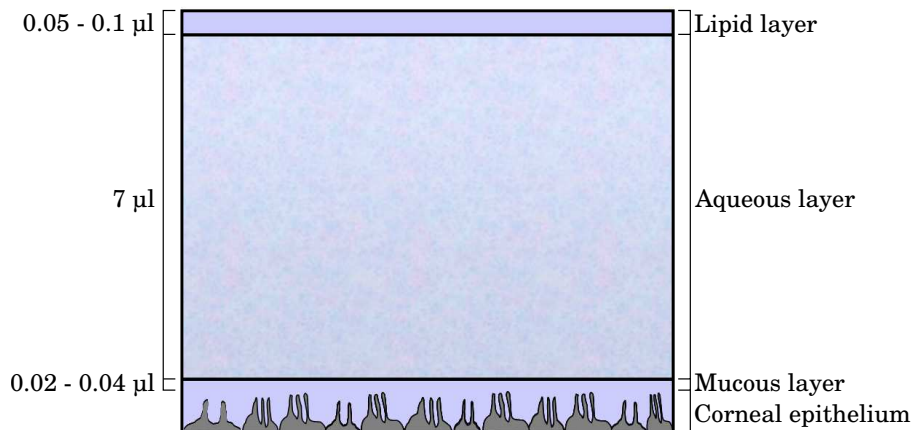


Figure 1.2: Structure of the tear film with the thickness and name of each layer.

1.1.1 Lipid layer

The lipid layer ($0.05 - 0.1 \mu\text{l}$) (Korb, 2002) comprises polar and non-polar lipids. Its main function is the reduction of evaporation from the aqueous phase. Moreover, its structure is important in preventing surface contamination which could disrupt the tear film. For this reason, it is the focus of various interferential techniques for tear film assessment.

1.1.2 Aqueous layer

The aqueous layer ($7\mu\text{l}$) (Korb, 2002; Larke, 1997) is the major component of the tear film, and represents around a 98% of its total thickness. This phase provides the proper functions of the tear film, and is mainly formed by proteins, metabolites, electrolytes and enzymes.

1.1.3 Mucous layer

The mucous layer ($0.02 - 0.04\mu\text{l}$) (Korb, 2002) is mainly formed by glycoproteins to maintain the corneal and conjunctival surfaces hydrated. The main function of these mucous glycoproteins is to reduce the surface tension of tears. Lubrication of the cornea is also an important function, since it allows the lids to smoothly slide with minimal friction during the blinking (Larke, 1997).

1.2 Dry eye syndrome

The *international dry eye workshop* (DEWS) established the main characteristics of the *dry eye syndrome* (DES) and published its finest definition (Lemp et al., 2007a) in 2007: “Dry eye is a multifactorial disease of the tears and ocular surface that results in symptoms of discomfort, visual disturbance and tear film instability with potential damage to the ocular surface. It is accompanied by increased osmolarity of the tear film and inflammation of the ocular surface.”

DES can be considered as an inflammatory status of the ocular surface driven by increased tear film osmolarity and derived by poor quality/quantity of the tear film. This disease affects quality of life, provokes visual disturbance and can lead to damage of the ocular surface.

DES has increased its prevalence in the last few years, reaching from 10 to 35% of the general population. If only contact lens wearers are considered, this prevalence is even greater (Lemp et al., 2007a). Current style of life, harmful environments such as pollution, tasks that favors increased tear film evaporation, and the aging of population have increased DES prevalence. For this reason, DES is currently considered an endemic condition.

1.2.1 The classification of dry eye syndrome

Two main categories of DES were defined by the DEWS (Lemp et al., 2007a) based on the main etiological causes of the disease. Figure 1.3 illustrates a contemporary

understanding of dry eye, which can be useful in order to apply different therapies according to severity of the disease. These two main categories are:

Aqueous tear-deficient dry eye (ADDE). This type of DES is age-related, and is produced by a failure of lacrimal tear secretion. Although it is mainly derived from a Sjogren syndrome, there are other factors which can occasion ADDE such as lacrimal gland infiltration, sarcoidosis, lymphoma, obstruction of the lacrimal gland ducts or reflex hyposalivation.

Evaporative dry eye (EDE). This type of DES refers to a normal lacrimal secretory function, and the tear film deficit is due to an excessive water loss from the exposed ocular surface. This is the type of dry eye most commonly found in young to middle-aged people, and related to ambient conditions such as air conditioning, and/or contact lens wear. EDE may be *intrinsic*, where the regulation of evaporative loss from the tear film is directly affected; or *extrinsic*, where it embraces those etiologies which increase evaporation by their pathological effects on ocular surface. The boundary between these two categories is inevitably blurred, although their characteristics are relevant for treatment and therapeutic protocols.

Notice that any form of dry eye can interact with and exacerbate other forms of dry eye, as part of a vicious circle.

Evaporation of the tear film in EDE

For any type of DES, hyperosmolarity is a precipitating event leading to the pathological changes associated with dry eye. In EDE, the rate of evaporation which results in critical osmolarity will depend on the tear flow rate. Evaporation rate is influenced by six different factors: ambient conditions, hormonal regulation, blink rate, area of palpebral aperture, tear film compartments, and tear film lipid layer (Foulks, 2007).

The outermost layer of the tear film, the tear film lipid layer, is a combination of polar and nonpolar lipids that are the secretion of the meibomian glands. As commented above, the chief function of the lipid layer is to retard water evaporation from the surface of the open eye (Foulks, 2007). In the normal tear film, much of the lipid layer is a structure that remains stable over a series of blinks, as it approaches the lower lid margin in the down-phase of the blink and unfolding in the up-phase, with little mixing of lipid within the lipid layer or between the lipid layer and the reservoirs (Bron, Tiffany, Gouveia, Yokoi, & Voon, 2004).

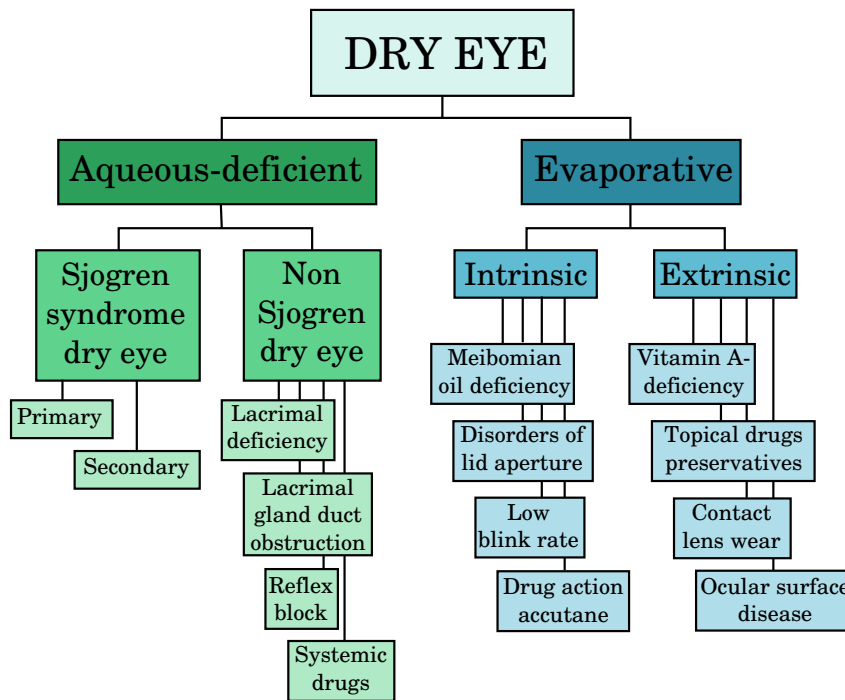


Figure 1.3: Major etiological causes of dry eye.

A stable tear film is one in which a minimum amount of tears evaporate. The evaporation rate is determined primarily by the status of the lipid layer, the protein constituents, aqueous components and the mucin coating the corneal epithelium (Foulks, 2007).

There is some evidence that evaporation is affected by lipid layer thickness, but it is currently not known specifically how lipid composition alters either the stability or thickness of the lipid layer (Bron & Tiffany, 2004). It has been proposed that the polar lipids act as a surfactant which helps spread the nonpolar lipids over the aqueous component of the tear film, provide a barrier between the two layers and also a structure that supports the nonpolar phase, which is responsible for creating a seal that decreases evaporation from the tear film (Foulks, 2007).

1.2.2 The epidemiology of dry eye syndrome

Epidemiology can be defined as a biomedical area that involves the research about the distribution of health and/or disease in human populations. In this manner, epidemiological studies allow the identification of the frequencies and the types of a particular disease, and the factors that influence its distribution.

Prevalence of dry eye syndrome

Dry eye syndrome is a common and frequently distressing condition. It affects a relatively large proportion of the population. Particularly, it has been estimated that about 3.23 million women and 1.68 million men, for a total of 4.91 million American people 50 years and older, have dry eye (Christen et al., 1998; Schaumberg, Sullivan, Buring, & Dana, 2003). Furthermore, tens of millions more have less severe symptoms which are more noticeable during contact with some adverse factors, such as low humidity or contact lens wear. Studies of age-specific data on the prevalence of the disease reveals that over 14% of 65+ age group in one US study (Moss, 2000), and over 30% of the same age group in a population of Chinese subjects (Jie, Xu, Wu, & Jonas, 2008) suffer from dry eye. The percentage of European people affected by dry eye is quite similar. In Germany, for example, one in four patients consulting an ophthalmologist complains of the symptoms of dry eye (Brewitt & Sistani, 2001). An overall summary of data from large epidemiological studies suggests that the prevalence of dry eye is in the range 5-35% at various ages.

Financial costs of dry eye syndrome

The high prevalence of dry eye among the older age groups, combined with the aging of the population, makes relevant the economic impact of the dry eye. Although few data exist on the direct and indirect costs of dry eye, it is well-known that many sufferers will require treatment and the potential cost is significant (Smith, 2007). The cost includes clinical visits, medicines and even surgery. In addition to the pain caused by the syndrome, intangible costs should be highlighted, such as impact in social interactions, decreased leisure time, and impaired quality of life. For all these reasons, monitoring the effect of the different treatments is of great importance in ensuring the maximum benefit to each individual (K. K. Nichols, Nichols, & Zadnik, 2000; Bron, 2001).

Impact of dry eye syndrome on quality of life

The dry eye syndrome affects the patients' quality of life in these main aspects (Lemp et al., 2007b): pain and irritating symptoms, effect on ocular and general health, effect on perception of visual function, and impact on visual performance. Also, dry eye limits performance of common daily activities, such as driving or working with computers (Schiffman, Christianson, Jacobsen, Hirsch, & Reis, 2000). The above mentioned cost of treatment and the lack of cure for dry eye add to the impact of this important public health problem (Lemp et al., 2007b).

1.3 Clinical tests for dry eye diagnosis

DES is a multifactorial syndrome, so several tests are necessary in order to obtain a clear diagnosis. There are a wide number of tests to evaluate different aspects of the tear film which can be divided into two main groups, depending on which tear film parameters they measured. On the one hand, quantitative tear film tests are related with the lacrimal gland secretion function and assess tear film tear secretion. On the other hand, qualitative tear film tests reflect the ability of the tear film to remains stable, which is essential to cover the anterior eye and perform its functions.

1.3.1 Quantitative tear film tests

These clinical tests assess the tear secretion, and the most common ones are:

Schirmer test. It is a test of reflex tear secretion in response to conjunctival stimulation (Schirmer, 1903). It is a useful test for the evaluation of dry eye, but the diagnosis cannot be made on the basis of this test alone. Also, it is the simplest test for assessing aqueous production by placing a blotting paper over the lower eyelid. See Figure 1.4a.

Phenol red thread test. It provides an index of tear volume, which is related to tear secretory rate and so detects aqueous-deficient dry eye syndrome. This test uses a cotton thread which has been treated with phenol red, a pH sensitive substance which changes from yellow to red in contact with the near neutral pH of the tears (Tomlinson, Blades, & Pearce, 2001). Note that the end of the cotton thread is gently placed over the lower eyelid, as in the Schirmer test. See Figure 1.4b.

Tear meniscus height. Tear meniscus volume is reduced in aqueous-deficient dry eye, as indicated by a reduced height and radius of curvature. In this sense, this clinical test measures the tear reservoir along the low lid, which is an indicator of tear volume. It is not invasive and only needs the observation of the tear meniscus by a slit-lamp (García-Resúa, Santodomingo-Rubido, Lira, Giráldez, & Yebra-Pimentel, 2009), which can be also observed by other optical devices. See Figure 1.5.

1.3.2 Qualitative tear film tests

These clinical tests assess the tear film stability, and the most common ones are:



Figure 1.4: (a) Schirmer test consists in placing a special paper over the lower eyelid to measure the tear production, whilst (b) phenol red thread test is similar but a cotton thread is placed under the lower eyelid.



Figure 1.5: (a) Appearance of tear meniscus height by slit-lamp biomicroscope. (b) Appearance of tear meniscus height by Tearscope Plus.

Lipid layer pattern assessment. Tear film quality and lipid layer thickness can be assessed by non-invasively imaging the superficial lipid layer with interferometry. The Tearscope Plus (Tearscope Plus, 1997) is the instrument of choice for rapid assessment of lipid layer thickness, and allows the qualitative analysis of the lipid layer structure. See Figure 1.6.

Tear break-up time (BUT). It is the standard clinical test for tear film stability (Lemp & Hamil, 1973), and it is considered an invasive test since the instillation of fluorescein needed shortens the normal break-up time. The break-up time is defined as the time that elapses from the last blink to the first appearance of a dark spot in the fluorescein-stained film. See Figure 1.7.

Not invasive tear break-up time (NIBUT). It is a non-invasive test of tear film stability which does not involve the instillation of fluorescein dye. The break-up time is measured as the time between the last blink and the break-up of a reflected image of a target on the tear film (Mengher, Bron, Tonge, & Gilbert, 1985). See Figure 1.8.

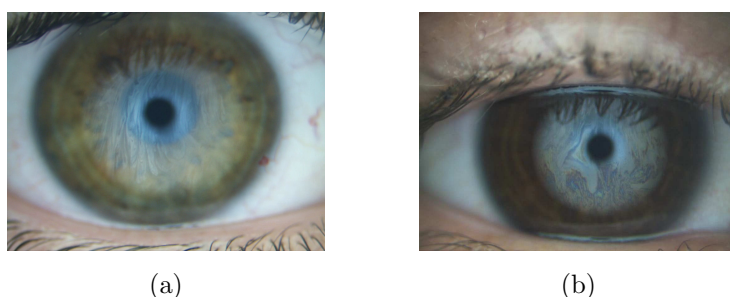


Figure 1.6: Appearance of the superficial lipid layer by interferometry.

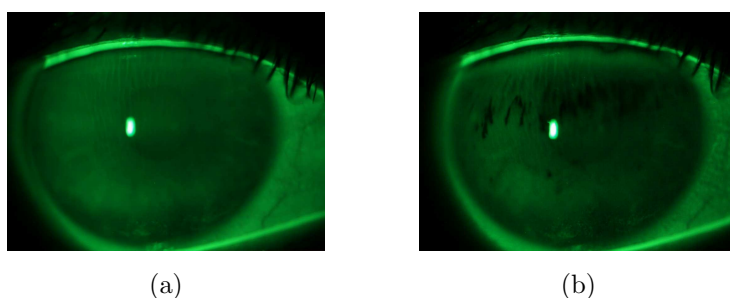


Figure 1.7: (a) Tear film stained by fluorescein. (b) Formation of dark spot points related to the tear film break-up.

This PhD thesis is focused on the test known as *lipid layer pattern assessment*, and so it will be subsequently explained in depth. The automated assessment of the lipid layer patterns is a first step in the path to developing a complete system, which will include the automation processes of the other clinical tests.

1.4 Lipid layer pattern assessment

The tear film is transparent, which makes difficult the direct observation during clinical assessments. For example, to assess whether tear film is present, BUT test requires staining the precocular tear film, whereas NIBUT test projects a grid on

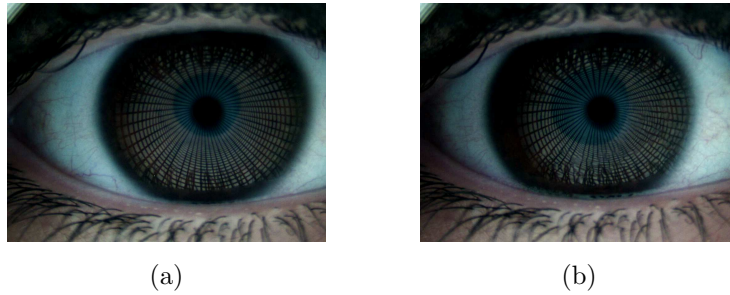


Figure 1.8: (a) Tear film after a blinking. (b) Grid deformation corresponding to the tear film break-up.

corneal surface. This problem is even greater when clinicians want to directly observe the component structures of the tear film. However, the structure of the tear film lipid layer can be appreciated *in vivo* by applying simple optic principles.

1.4.1 Specular reflection

Due to the incidence of a light source on an interface between two refractive index media, a small percentage of the incident light is specularly reflected. Because the refractive index of the lipid layer is higher than that of the aqueous layer, there is a second interface, between the two layers, which can be visible in specular observation. The observation of these specular reflections permits the evaluation of the precorneal tear film structure. This has been used to observe the anterior lipid layer with slit-lamp biomicroscope, but it only allows the observation of a $1\text{mm} \times 2\text{mm}$ area, because the light source of the biomicroscope subtends only a small angle. To solve this, McDonald (McDonald, 1969) introduced a hemispherical medical lamp to obtain large reflection by the tear film, so larger areas of superficial lipid layer can be evaluated, and posterior devices followed this design (see Figure 1.9).

1.4.2 Interference phenomena

When observing the appearance of the lipid layer by this technique, the presence of interference fringes can be appreciated. These interference fringes result from the wave characteristics of light, and the fact that when coherent rays of light of a given wavelength are combined and brought to a common focus, they will interfere, either constructively or destructively, depending on the degree to which the periodic fluctuations of their electromagnetic fields are in phase. To observe interference phenomena, it is necessary to use coherent light sources, i.e., sources whose phase difference remains constant in time. A simple manner in which this can be

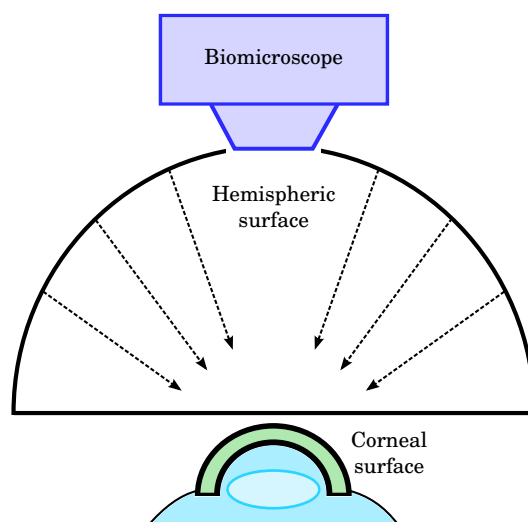


Figure 1.9: Hemispherical light source to obtain large reflection by the tear film.

accomplished is by using a single light source and its optical image.

In the case of the tear film, there are two interfering beams: the beam reflected from the air-lipid interface of the tear film, and the beam reflected from the lipid-aqueous interface of the tear film. These two beams originate from the same point of the single light source and, in fact, are two images of it, so the beams satisfy the requirement of coherence. Figure 1.10 shows an schema of this phenomena between two flat boundaries, air-lipid boundary and lipid-aqueous boundary.

This interference phenomena can be visible by the specular reflection commented above, and so the observer can appreciate an interference pattern. This pattern is formed by fringes and/or colors, and is commonly known as tear film lipid layer pattern. Color fringes are related with lipid layer thickness so the determination of lipid layer thickness can be extrapolated. However, the lipid reflection does not always show a color pattern. The observation of a colorless pattern (gray color) is because its thickness is below the minimal thickness to produce interference fringes. Korb (Korb, 2002) established the lipid layer thickness which corresponds to each color (see Table 1.1), by using a custom-designed hemicylindrical broad-spectrum illumination source and slit-lamp biomicroscope.

1.4.3 Tearscope Plus

Several devices, based on the optical principles previously exposed, have been designed to assess the lipid layer patterns through the interference phenomena. The

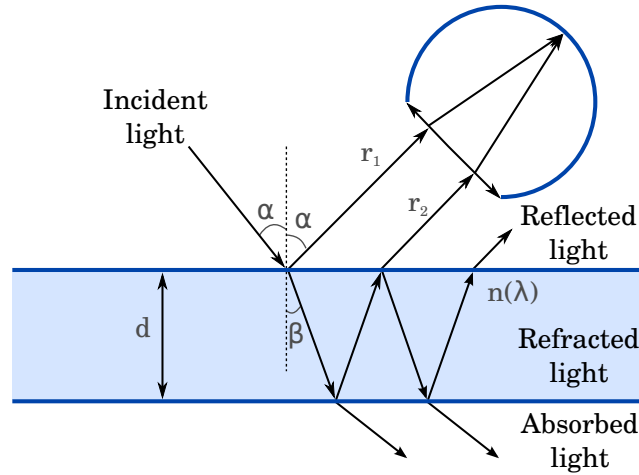


Figure 1.10: Optical diagram which shows the interference phenomena. Both light beams r_1 and r_2 are originate from the same source: r_1 is the light beam reflected from the air-lipid interface of the tear film, and r_2 is light beam reflected from the lipid-aqueous interface of the tear film. The thickness d generates an optical path difference between them, and will produce interference fringes after recombination; α is the incidence angle, equal to the reflected angles; β is the refracted angle; and $n(\lambda)$ is the refraction index.

Table 1.1: Color of the interference patterns and their lipid layer thickness.

	Lipid layer thickness (nm)
Grey to white	30-60
Grey/yellow	75
Yellow	90
Yellow/brown	105
Brown/yellow	120
Brown	135
Brown/blue	150
Blue/brown	165
Blue	180

Tearscope Plus is the instrument employed by the team from the Faculty of Optics and Optometry (University of Santiago de Compostela, Spain) who have collaborated in this research.

The Tearscope Plus (Tearscope Plus, 1997) was designed by Guillon as an instrument for the non-invasive examination of the tear film, its appearance, volume, stability, and its effect on the ocular and contact lens surface. It is a hand-held instrument which can be used alone or in conjunction with a biomicroscope (Guillon, 1998) (see Figure 1.11). The first way makes faster the lipid layer pattern evaluation, although is recommended to use with the biomicroscope to obtain images with high magnification. The Tearscope Plus projects a cylindrical source of cool white fluorescent light onto the lipid layer. Thus, any observed phenomena is unique to the specific light source of this device. The Tearscope Plus lighting system is a diffuse hemispherical light source with a central hole to allow observation (Guillon, 1998).

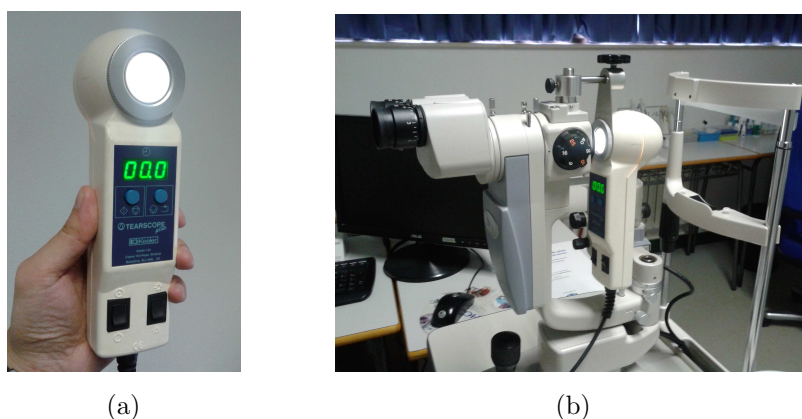

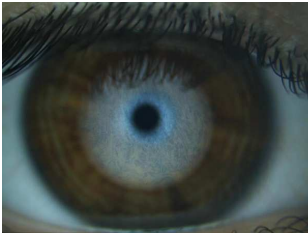

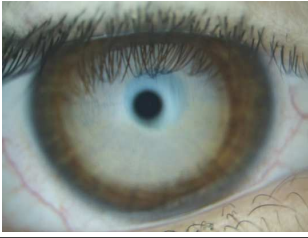
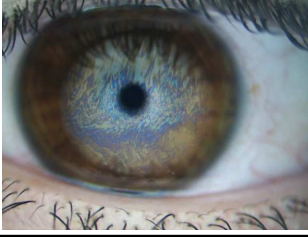


Figure 1.11: (a) Tearscope Plus hand-held instrument. (b) Tearscope Plus attached to a slit-lamp for high magnification.

The grading of the lipid layer appearance in its undisturbed state should always be the first clinical observation to be made (Craig & Tomlinson, 1997). Practitioners need to recognize the different types of patterns: the pattern linked to the most stable tear film, as it represents the best candidate for comfortable contact lens wear; the pattern linked to increase evaporation and reduced stability; the normal pattern linked to average stability; and the pattern of thin coverage that may not form continuously over a contact lens. In order to facilitate this task, Guillon proposed five main grades of lipid layer thickness interference patterns for observations made using the Tearscope Plus (Guillon, 1998). These patterns are based on morphological, color features as it can be seen in Table 1.2.

Although this method offers a useful technique to evaluate the quality and struc-

Table 1.2: Appearance and estimated thickness of the tear film lipid layer patterns observed with the Tearscope Plus.

	Open meshwork	
	It represents a very thin, poor and minimal lipid layer stretched over the ocular surface. It is a gray, marble-like pattern, prone to evaporative dry eye.	~13-50 nm
	Closed meshwork	
	It indicates more lipid than open meshwork, less stretching of the lipid film. It is a gray, marble-like pattern, but with closed meshwork and tight pattern.	~13-50 nm
	Wave	
	It is thicker than meshwork with wavy, gray streak effect. This represents average tear film stability.	~50-70 nm
	Amorphous	
	It is associated with a thick, white yellowish even and well mixed lipid layer that may show colors during the blink. Ideal candidate for contact lens fitting.	~80-90 nm
	Color fringe	
	It is a thicker lipid layer with mix of brown and blue fringes. Good candidate for contact lens wear with possible tendency for greasing problems.	~90-180 nm

ture of the tear film, it is affected by the subjective interpretation of the observer. Thick lipid layers ($\geq 90nm$) are readily observed since they produce color and wave patterns. However, thin lipid layers ($\leq 60nm$) are difficult to observe, since color fringes and other distinct morphological features are not present, so visualizations may be affected by the subjective interpretation of the observer (Korb, 2002). Training also affects the interpretation of the patterns according to the *learning curve* established for lipid layer pattern grading (J. J. Nichols, Puent, Saracino, & Mitchell, 2002). The Tearscope Plus can be also used with a camera attached to a slit-lamp (García-Resúa et al., 2013), so lipid layer videos can be stored for further analysis.

1.5 Image datasets

The procedure for image acquisition, and the different image datasets used in this research are subsequently described. These datasets were acquired in different illumination conditions, and annotated by different optometrists in order to test the proposed automated assessments. All images have been acquired and annotated by optometrists from the Faculty of Optics and Optometry, University of Santiago de Compostela (Spain).

1.5.1 Image acquisition

The input image acquisition was carried out with the Tearscope Plus (Tearscope Plus, 1997) attached to a Topcon SL-D4 slit-lamp (Topcon SL-D4, n.d.). The Tearscope Plus was designed by Guillon (Guillon, 1998) as an instrument for rapid evaluation of the lipid layer thickness in clinical settings. This instrument projects a cylindrical source of cool white fluorescent light onto the lipid layer illuminating almost all of the corneal surface area. The interference patterns were observed through a slit-lamp biomicroscope, with magnification set at 200X.

The Tearscope Plus is attached to a slit-lamp in the image acquisition procedure: the lipid layer is focused with the slit-lamp, and then the Tearscope Plus is approached toward the patient's eye. The closer the Tearscope Plus to the subject, the higher the lipid layer area. It has been figured out that lipid layer patterns are more difficult to categorize in clear eyes than in dark eyes, because in the former ones the iris features could be seen through the lipid layer, which could diminish the visibility of the pattern. In order to avoid this problem, the lipid layer area has to be reduced, so the intensity of the lipid layer pattern would be more concentrated in this area and the iris features would be less visible. For this reason, two sizes of areas were considered: the biggest one, when no iris features were visible through

the lipid layer (see Figure 1.12a); and the smallest one, where the area was halved (see Figure 1.12b). In the acquisition procedure all patterns were initially acquired with the biggest area, but in those cases in which the iris was visible, the smallest area was used.

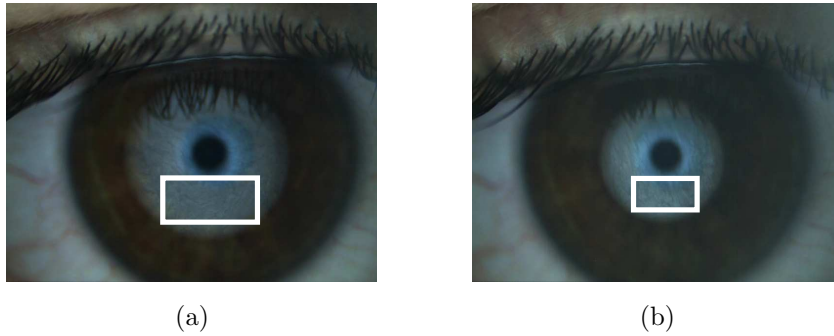


Figure 1.12: (a) Pattern with the biggest area. (b) Pattern with the smallest area.

Since the tear film is not static between blinks, a video was recorded and analyzed by an optometrist in order to select the best images for processing. Those images are exactly the same that the specialists analyzed by hand. In order to get them, the interference phenomena was recorded using a Topcon DV-3 digital video camera (Topcon DV-3, n.d.), and stored at a computer via the Topcon IMAGEnet i-base (Topcon IMAGEnet, n.d.). Then, an image was selected to go through the processing step only when the tear film was completely expanded after the eye blink, according to the expert's criterion. Note that the images have a spatial resolution of 1024×768 pixels per frame in the RGB color space.

1.5.2 Illumination conditions

Although the interference patterns are independent of the illumination, there is an optimum range of illuminations used by optometrists to obtain the images. Images with illuminations outside this range are considered noisy images. Figure 1.13 shows an example of two images from the same subject. It can be seen that a too high illumination produces an image where the interference pattern is hardly appreciated.

1.5.3 VOPTICAL_I1 dataset

The VOPTICAL_I1 dataset (VOPTICAL_I1, n.d.) contains 105 images of the pre-ocular tear film taken over optimum illumination conditions. These images were acquired from healthy patients with ages ranging from 19 to 33 years. The dataset

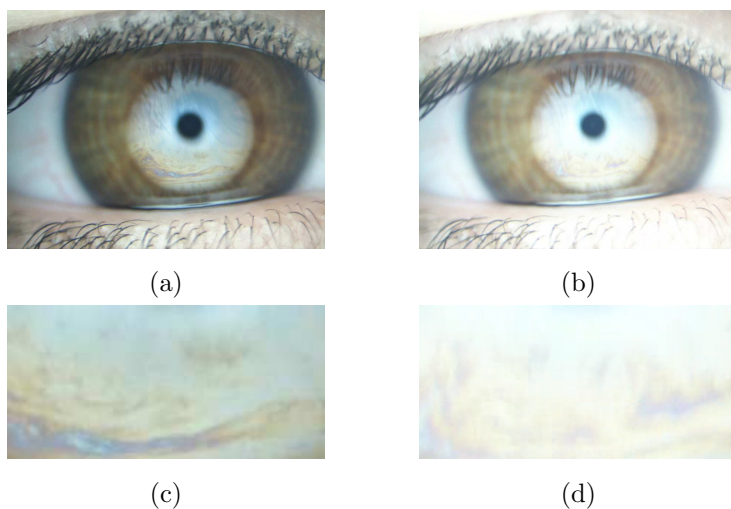


Figure 1.13: (a) Image obtained using an optimum illumination, and (c) its central area in which a color fringe pattern is clearly observable. (b) Image obtained using a too high illumination, and (d) its central area in which a color fringe pattern is hardly observable.

includes 29 open meshwork, 29 closed meshwork, 25 wave and 22 color fringe images. The annotation of each single image is one of the four Guillon categories considered (open meshwork, closed meshwork, wave and color fringe).

The images of this dataset were taken over the same illumination conditions, which are considered to be the optimum ones by practitioners. This dataset contains the samples that are expected to be obtained in a real case situation, and will be used to compute the performance of algorithms.

1.5.4 VOPTICAL_Is dataset

The VOPTICAL_Is dataset (VOPTICAL_Is, n.d.) contains 406 images of the pre-ocular tear film taken over four different illuminations. These images were acquired from healthy patients with ages ranging from 19 to 33 years. The dataset includes 159 open meshwork, 117 closed meshwork, 90 wave and 40 color fringe images. The annotation of each single image is one of the four Guillon categories considered (open meshwork, closed meshwork, wave and color fringe).

The images of this dataset were taken over different illumination conditions. This bank will be used only in some specific experiments, in order to evaluate the sensibility of algorithms to *noisy* data.

1.5.5 VOPTICAL_R dataset

The VOPTICAL_R dataset contains 44 images of the preocular tear film taken over optimum illumination conditions. These images were acquired from healthy patients with ages ranging from 19 to 33 years. The annotations consist of delimited regions in the images associated with the five Guillon categories (open meshwork, closed meshwork, wave, amorphous and color fringe). Each expert has annotated an average of 87 regions over the 44 images.

The images of this dataset were taken over the same illumination conditions, which are considered to be the optimum ones by practitioners. This dataset contains the samples that are expected to be obtained in a real case situation, and will be used to compute the performance of algorithms.

1.6 Thesis

The lipid layer thickness can be evaluated by the classification of the interference patterns. In this sense, the Tearscope Plus is the instrument designed by Guillon for rapid assessment of lipid layer thickness (Guillon, 1998). Another devices were designed for lipid layer examination, but the Tearscope Plus is still the most commonly used instrument in clinical settings and research.

Guillon defined five main grades of lipid layer patterns (Guillon, 1998) to evaluate the lipid layer thickness through the Tearscope Plus. However, the classification into these grades is a difficult clinical task, especially with thinner lipid layers that lack color and/or morphological features. The subjective interpretation of the experts via visual inspection may affect the classification. This time-consuming task is very dependent on the training and experience of the optometrist(s), and so produces a high degree of inter- and also intra- observer variability (García-Resúa et al., 2013). The development of a systematic, objective computerized method for analysis and classification is thus highly desirable, allowing for homogeneous diagnosis and relieving the experts from this tedious task.

The proposal of this research is to design an automatic system to perform different assessments of the tear film lipid layer patterns. This system is based on the interpretation of the images acquired with the Tearscope Plus, and on the five categories defined by Guillon. Different image processing techniques and machine learning algorithms are applied in the development and validation of the automated assessments following presented.

Chapter 2 describes the methodology to assess the tear film lipid layer by automatically classifying images acquired with the Tearscope Plus into the Guillon

categories. The process is carried out using texture and color analysis techniques, and machine learning algorithms.

The previous approach provides results at the expense of a too long processing time and too much memory, since many features have to be computed. This fact makes this methodology unfeasible for practical applications and prevents its clinical use. The reduction of the computational complexity of the previous approach is tackled in Chapter 3 by applying dimensionality reduction methods. This optimization is focused not only on the improvement of the accuracy, but also on the reduction of both memory and time requirements.

Since the heterogeneity of the tear film lipid layer makes its classification into a single category not always possible, tear film maps are presented in Chapter 4 in order to illustrate the local distribution of the lipid layer patterns. In this manner, more memory and time requirements are needed in exchange for a more detailed information about the localization and size of the patterns over the tear film.

Finally, Chapter 5 provides a brief overview of some concluding remarks and proposes some future lines of research.

Chapter 2

Tear film assessment

Optometrists carry out tear film assessment by means of the evaluation of the lipid layer through a manual process, which consists in classifying images obtained with the Tearscope Plus into one of the Guillon categories. The Tearscope Plus has proven its validity to the lipid layer pattern assessment (Rolando, Valente, & Barabino, 2008; García-Resúa et al., 2013). However, many eye care professionals have abandoned this test because the difficulty interpreting the lipid layer patterns, specially the thinner ones; and the lack of a huge bank of images for reference purposes. Nevertheless, there is no doubt that the examination of the structure of the tear film lipid layer is a valuable technique which provides practitioners with relevant information about the stability of the tear film by using noninvasive procedures.

This clinical task is not only difficult and time-consuming, but also affected by the subjective interpretation of the observers. This has motivated the development of automated techniques to characterize the interference phenomena characteristic of the lipid layer patterns, in such a way that the tear film lipid layer can be automatically classified into one of the categories enumerated by Guillon. Thus, this chapter presents a research methodology which, from a photography of the eye, detects a region of interest and extracts its low-level features, generating a feature vector which describes it, to be finally classified into one of the target categories.

Next section presents the proposed methodology to automatically assess the tear film by interference phenomena. Following, its steps will be explained in depth. Firstly, the procedure to locate the region of interest of a single image, in which the analysis will take place, is explained. Secondly, different color spaces and texture analysis methods are proposed to compute the low-level features of the images. Next, several machine learning algorithms are described in order to classify the final feature vector into one of the Guillon categories. Finally, the conclusions of the proposed methodology for tear film assessment are briefly exposed and discussed.

2.1 Research methodology

The proposed methodology is composed of three main steps (see Figure 2.1): from an input image acquired with the Tearscope Plus, its region of interest (ROI) is located and some low-level features are extracted from it, and finally the image is classified into one of the Guillon categories.

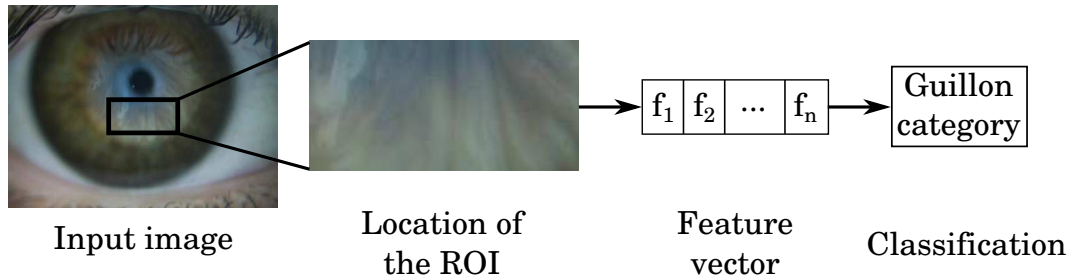


Figure 2.1: Steps of the research methodology to assess tear film lipid layer patterns.

These three steps will be subsequently presented in depth, including the experimentation performed in each of them. Roughly speaking, the steps are as follows:

1. *Location of the region of interest.* This stage aims at finding the area of the input image where the tear film can be observed with higher contrast. This area will correspond to the so-called region of interest, where the following analysis will take place.
2. *Feature vector.* The low-level features of the region of interest are extracted in this step. Color and texture are the two discriminant features of the Guillon categories: thicker lipid layers show defined patterns while thinner layers are more homogeneous, and some categories show distinctive color characteristics.
3. *Classification.* The last stage classifies an input image into one of the Guillon categories using its feature vector and a machine learning algorithm. This algorithm will be able to learn based on the training data, and so it could make predictions in the future.

2.2 Location of the region of interest

The input images, as depicted in Figure 2.2, include several areas of the eye which do not contain relevant information for the classification, such as the sclera, eyelids and eyelashes. Optometrists that analyze these images usually focus their attention

on the bottom part of the iris, since this is the area in which the tear can be perceived with better contrast. This forces a preprocessing step aimed at extracting the region in which the tear film classification takes place, called *region of interest* (ROI) (Calvo, Mosquera, Penas, García-Resúa, & Remeseiro, 2010).

Figure 2.2 depicts an example of the process performed to locate the ROI. The acquisition procedure generates a central area in the image, more illuminated than the others. This area corresponds to the region used by the optometrists to assess the tear film by interference phenomena and, thus, to the ROI. As the illumination plays an essential role, the input image in RGB is transformed to the Lab color space and only its component of luminance L is selected in this stage. Then, the normalized cross-correlation between the L component of the image and a set of ring-shaped templates previously generated, that cover the different ROI shapes, is computed. Next, the region with maximum cross-correlation value is selected and, as the region of interest is situated at the bottom part, the top area is rejected. Finally, the rectangle of maximum area inside this bottom part is located and so the ROI of the input image is obtained through a completely automatic process.

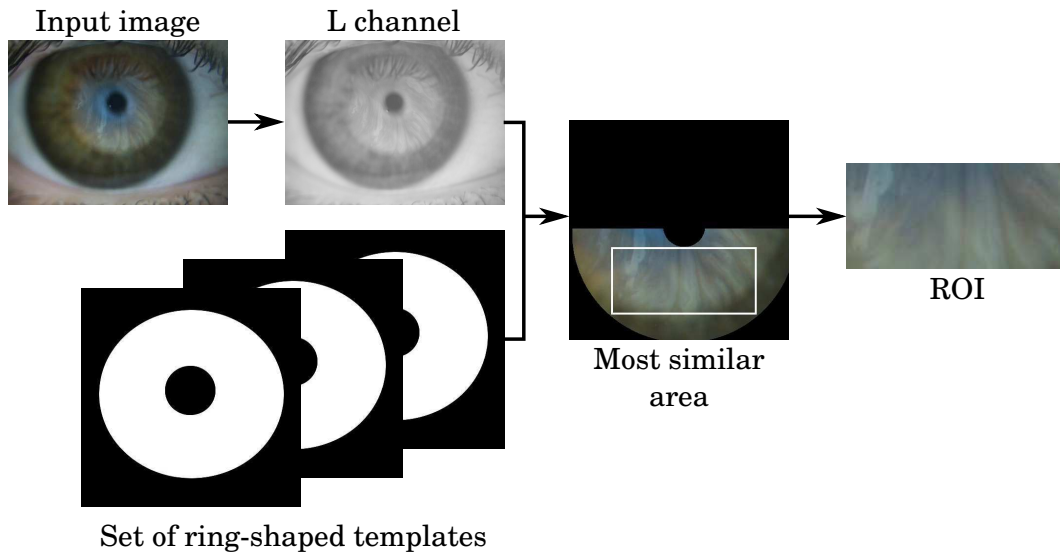


Figure 2.2: Location of the region of interest over a representative image.

2.2.1 Experimental study

The objective is to check the behavior of the procedure to locate the ROI over the Tearscope images, and so an experiment was performed to qualitatively analyze it.

The experimental procedure consists in applying the process to locate the ROI over all the images of the VOPTICAL_I1 dataset. Next, the effectiveness of the method was quantitatively evaluated by means of a visual inspection, which determines if the ROI is properly located at the bottom part of iris or not.

The obtained results indicate that the location of the ROI is appropriate in all the images tested. As an example, the ROIs of three representative images are illustrated in Figure 2.3. Notice that these input images correspond to different situations since, for example, images (a) and (b) have bigger ROIs than image (c).

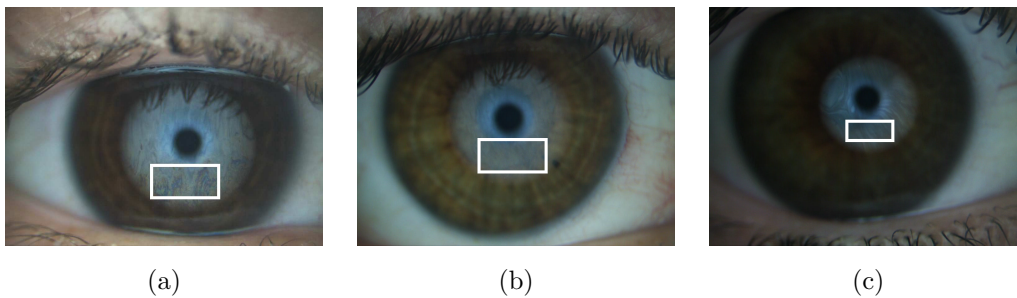


Figure 2.3: Representative input images and location of their ROIs.

2.3 Feature vector

After extracting the region of interest, the next step entails analyzing its low-level features by means of color and texture analysis. Color and interference patterns are the two discriminant features of the Guillon categories for tear film classification. On the one hand, some categories show distinctive color characteristic which motivates the color analysis step. On the other hand, the interference phenomena can be characterized as a texture pattern, since thicker lipid layers show defined patterns while thinner layers are more homogeneous.

2.3.1 Color analysis

Color is one of the discriminant features of the Guillon categories. For this reason, the color present in the Tearscope images is represented by means of two different color models (Ramos et al., 2011). On the one hand, the RGB color model which is based on the physiology of the human eye. It is composed of three colors which fall within each of the sensitivity ranges of each of the human cone photoreceptors. Since it is not perceptually uniform, this color model is used in this research through the opponent color theory. On the other hand, the Lab color model which is based

on the color perception of the human brain. This color space is defined by three components: one represents the perception of the illumination, and the other two represent the perception of the tone and saturation, i.e., the chromaticism. In addition, the grayscale images are also considered in order to verify the appropriate consideration of using color information. These three options for color analysis are subsequently explained.

Grayscale

A grayscale image is one in which the only color is gray, represented by different levels from black to white. In this case, less information needs to be provided since it is only necessary to specify a single intensity value for each pixel.

In order to generate a grayscale image, the three channels of the RGB image (R , G and B) have to be converted into only one gray channel (Gr), according to the following expression (Bradski, 2000):

$$Gr = 0.299 \cdot R + 0.587 \cdot G + 0.114 \cdot B \quad (2.1)$$

The RGB color space: opponent colors

The RGB color space (Sangwine & Horne, 1998) (RGB) is an additive color space based on the physiology of the eye. It is defined by three chromatic components: the red channel R , the green channel G , and the blue channel B . Despite being one of the most frequently used color spaces for image processing, it is not perceptually uniform. Therefore, the opponent process theory of human color vision, proposed by Hering (Hering, 1964) in the 1800s, is considered. This theory states that the human visual system interprets information about color by processing three opponent channels: red vs. green (R_G), green vs. red (G_R) and blue vs. yellow (B_Y). The three opponent channels have to be calculated from the RGB image according to (Borer & Süssstrunk, 2002):

$$\begin{aligned} R_G &= R - p * G \\ G_R &= G - p * R \\ B_Y &= B - p * (R + G) \end{aligned} \quad (2.2)$$

where p is a low pass filter.

The Lab color space

The CIE 1976 L*a*b color space (McLaren, 1976) (Lab) is a chromatic color space which describes all the colors that the human eye can perceive. It was defined by

the International Commission on Illumination, abbreviated as CIE from its French title *Commission Internationale de l'Eclairage*. Lab is a 3D model where its three coordinates represent: the luminance of the color L , its position between magenta and green a , and its position between yellow and blue b . Its use is recommended by CIE in images with natural illumination. In addition, this color space is perceptually uniform, which means that a change of the same amount in a color value produces a change of the same visual importance. This characteristic is also important since the specialists' perception is being imitated.

The use of the Lab color space entails converting the three channels of the RGB image into the three components of Lab. This transformation has to be done by using the CIE XYZ color space and its three channels X , Y and Z (Bradski, 2000):

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} 0.4124563 & 0.357580 & 0.180423 \\ 0.212671 & 0.715160 & 0.072169 \\ 0.019334 & 0.119193 & 0.950227 \end{bmatrix} \cdot \begin{bmatrix} R \\ G \\ B \end{bmatrix} \quad (2.3)$$

$$\begin{aligned} X &= X/0.950456 \\ Z &= Z/1.088754 \end{aligned} \quad (2.4)$$

Next, the Lab channels are calculated according to:

$$L = \begin{cases} 116 \cdot Y^{1/3} - 16 & \text{for } Y > 0.008856 \\ 903.3 \cdot Y & \text{for } Y \leq 0.008856 \end{cases} \quad (2.5)$$

$$\begin{aligned} a &= 500(f(X) - f(Y)) + 128 \\ b &= 200(f(Y) - f(Z)) + 128 \end{aligned} \quad (2.6)$$

where:

$$f(t) = \begin{cases} t^{1/3} & \text{for } t > 0.008856 \\ 7.787t + 16/116 & \text{for } t \leq 0.008856 \end{cases} \quad (2.7)$$

2.3.2 Texture analysis

Texture is used to characterize the interference patterns of the five categories defined by Guillon (Remeseiro et al., 2011) (see Figure 2.4). Several techniques for texture analysis could be applied and, in this study, five popular methods were tested: Butterworth filters, Gabor filters and the discrete transform as signal processing methods; Markov random fields as a model based method; and co-occurrence features as an statistical method. All these methods are subsequently described.

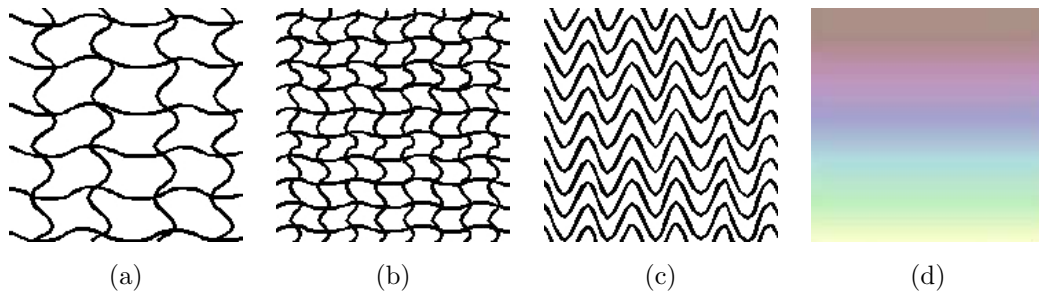


Figure 2.4: Simplified texture patterns: (a) open meshwork, (b) closed meshwork, (c) wave, (d) color fringe.

Butterworth filters

Butterworth band-pass filters (Gonzalez & Woods, 2008) are frequency domain filters that have a flat response in the band-pass frequency, which gradually decays in the stopband. A Butterworth filter can be represented in 1D as:

$$f(\omega) = \frac{1}{1 + \left(\frac{\omega - \omega_c}{\omega_0}\right)^{2n}} \quad (2.8)$$

where n is the order of the filter, ω the angular frequency, ω_0 the cutoff frequency and ω_c the center frequency. The order n of the filter defines the slope of the decay: the higher the order, the faster the decay.

A bank of 9 second order filters is used, so that the whole frequency spectrum is covered by the band-pass frequencies considered. By that means, the filter bank maps each input image into 9 filtered images, one per frequency band.

The results of each frequency band have to be normalized, and the histograms of their output images have to be computed. Analyzing those histograms, it can be seen that they concentrated most of the information in the lower bins, which made their comparison difficult. For this reason, histograms with equiprobable bins, i.e., with non-equidistant bins, are computed instead of the traditional ones. The process to obtain uniform histograms is described as follows: given all the filtered images of an specific frequency band, the limits of the histogram are defined so that each bin contains a maximum of $\frac{N}{N_{bins}}$ pixels, where N is the number of pixels in the corresponding frequency and N_{bins} the number of histogram bins.

Using 16-bin histograms, the descriptor of an input image has 16 components per frequency band.

Gabor filters

Gabor filters (Gabor, 1946) are complex exponential signal modulated by Gaussian functions widely used in texture analysis. A 2D Gabor filter (Daugman, 1985), using cartesian coordinates in the spatial domain and polar coordinates in the frequency domain, can be defined as:

$$g_{x,y,f,\theta} = \exp\{i[2\pi f(x\cos\theta + y\sin\theta) + \phi]\} \text{gauss}(x, y) \quad (2.9)$$

where

$$\text{gauss}(x, y) = a \cdot \exp\{-\pi [a^2(x\cos\theta + y\sin\theta)^2 + b^2(x\sin\theta - y\cos\theta)^2]\} \quad (2.10)$$

a and b model the shape of the filter; while x , y , f and θ represent the location in the spatial and frequency domains, respectively.

A bank of filters is created with 16 Gabor filters centered at 4 frequencies and 4 orientations. Thus, the filter bank maps each input image to 16 filtered images, one per frequency-orientation pair.

Using the same idea as in Butterworth filters, the feature vector is created by generating the uniform histogram with non-equidistant bins.

The discrete wavelet transform

The discrete wavelet transform (Mallat, 1989) generates a set of wavelets by scaling and translating a *mother wavelet*, which is a function defined both in the spatial and frequency domain, that can be represented in 2D as:

$$\phi^{a,b}(x, y) = \frac{1}{\sqrt{a_x a_y}} \phi\left(\frac{x - b_x}{a_x}, \frac{y - b_y}{a_y}\right) \quad (2.11)$$

where $a = (a_x, a_y)$ governs the scale and $b = (b_x, b_y)$ the translation of the function. The values of a and b control the band-pass of the filter in order to generate high-pass (H) or low-pass (L) filters.

The wavelet decomposition of an image consists in applying wavelets horizontally and vertically in order to generate 4 subimages at each scale (LL, LH, HL and HH), which are then subsampled by a factor of 2. After the decomposition of the input image, the process is repeated $n - 1$ times over the LL subimage, where n is the number of scales of the method. This iterative process results in the so-called standard pyramidal wavelet decomposition.

Some statistical measures are used in order to create the descriptor from an input image: mean, absolute average deviation and energy. These measures are respectively defined as:

$$\mu = \frac{1}{N} \sum_{i=1}^N p(i) \quad (2.12)$$

$$aad = \frac{1}{N} \sum_{i=1}^N |p(i) - \mu| \quad (2.13)$$

$$e = \frac{1}{N^2} \sum_{i=1}^N p(i)^2 \quad (2.14)$$

where $p(i)$ is the i th entry in the image, and N represents its number of pixels.

The feature descriptor of an input image is constructed from the μ and the aad of the input and LL images, and from the e of the LH, HL and HH images.

Different mother wavelets can be considered, and the most popular ones are Haar and Daubechies (Daubechies, 1992). Haar is the simplest nontrivial wavelet and Daubechies is one representative type of basis for wavelets. Daubi represents the Daubechies orthonormal wavelet, where the number of vanishing moments is equal to half the coefficient i . Notice that the Haar wavelet is equivalent to Daub2.

Markov random fields

Markov random fields (MRF) (Besag, 1974) are model based texture analysis methods which construct an image model whose parameters capture the essential perceived qualities of the texture. A MRF model is based upon the assumption that a pixel intensity distribution is conditionally dependent upon only its local neighborhood, and independent of the rest of the image. Thus, MRFs generate a texture model by expressing the gray values of each pixel in an image as a function of the gray values in its neighborhood.

The concept of neighborhood is defined as the set of pixels within a distance d , and the Chebyshev distance is considered. The Markov process for textures is modeled using a *Gaussian Markov random field* (GMRF) defined as (Woods, 1972):

$$X(c) = \sum_m \beta_{c,m} [X(c+m) + X(c-m)] + e_c \quad (2.15)$$

where e_c is the zero mean Gaussian distributed noise, m is an offset from the center cell c , and $\beta_{c,m}$ are the parameters which weigh a pair of symmetric neighbors to

the center cell. The β coefficients describe the Markovian properties of the texture and the spatial interactions among pixels.

Equation (2.15) can be represented in matrix notation as:

$$X(c) = \beta^T Q_c + e_c \quad (2.16)$$

and, consequently, the β coefficients can be estimated through least squares fitting:

$$\beta = \left[\sum_{c \in I} Q_c Q_c^T \right]^{-1} \left[\sum_{c \in I} Q_c X(c) \right] \quad (2.17)$$

The descriptor of an input image is composed of the directional variances proposed by Çesmeli and Wang (Çesmeli & Wang, 2001), which are defined as:

$$f_i = \frac{1}{N \times M} \sum_{c \in I} [X(c) - \beta_i Q_{ci}]^2 \quad (2.18)$$

where $N \times M$ represents the dimensions of the input image.

For a distance d , the descriptor comprises $4d$ features. Different distances can be considered and their descriptors can be combined by means of their concatenation.

Co-occurrence features

Co-occurrence features analysis (Haralick, Shanmugam, & Dinstein, 1973) is a popular and effective texture descriptor based on the computation of the conditional joint probabilities of all pairwise combinations of gray levels, given an interpixel distance d and an orientation θ . The method generates a set of *gray level co-occurrence matrices* (GLCM), and extracts several statistical measures from their elements.

As in the above method, the *Chebyshev* distance is considered. For a distance $d = 1$, four orientations are considered (0° , 45° , 90° and 135°), and so four GLCMs are generated. In general, the number of orientations and, accordingly, the number of matrices for a distance d is $4d$.

From each GLCM, a set of 14 statistics proposed by Haralick et al. (Haralick et al., 1973) are computed. For explanatory purposes, the definition of 2 of these 14 statistical measures is shown:

$$f_1 = \sum_{i=1}^N \sum_{j=1}^N \left(\frac{P_{\theta,d}(i,j)}{R} \right)^2 \quad (2.19)$$

$$f_2 = \sum_{n=0}^{N-1} n^2 \left\{ \sum_{|i-j|=n} \left(\frac{P_{\theta,d}(i,j)}{R} \right) \right\} \quad (2.20)$$

where $P_{\theta,d}(i,j)$ are the elements of the GLCM, N is the number of distinct gray levels in the input image, and R is a normalizing constant. The angular second-moment feature f_1 is a measure of homogeneity of the image, and the contrast feature f_2 is a measure of the amount of local variations present in the image. Appendix B includes the definition of the whole set of measures.

Finally, the mean and the range of these 14 statistics are calculated across matrices and a set of 28 features composes the texture descriptor for a distance d .

2.3.3 Definition of the feature vector

The feature vector of an input image is created by using the color models and texture descriptors previously presented. The process is slightly different depending on the color analysis method considered. Regarding grayscale images, the gray channel obtained is analyzed in terms of texture and so the final descriptor is obtained (see Figure 2.5). However, the process changes using the RGB and Lab color spaces since they have three channels instead of one. In these two cases, each component is analyzed separately and its texture descriptor is obtained, so the final descriptor is the concatenation of the three descriptors (see Figures 2.7 and 2.6, respectively).

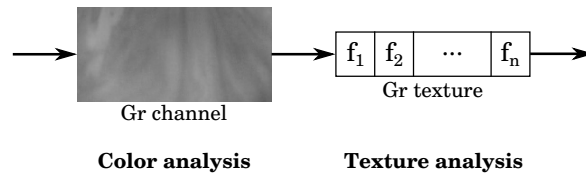


Figure 2.5: Feature vector steps using grayscale images.

Notice that when different filters or neighborhoods can be used in the corresponding methods, their individual descriptors can be combined by means of their concatenation, independently of the color analysis method.

2.3.4 Experimental study

The objective is to find which color and texture properties describe better the interference phenomena characteristic of the lipid layer patterns, and so they are the most appropriate for this problem. A total of 6 different experiments were carried out: one per each texture analysis method, and an extra experiment with all the possible combinations of methods.

The experimental procedure is detailed in Figure 2.8. Firstly, the three color analysis and the five texture analysis methods were applied to the VOPTICAL_I1

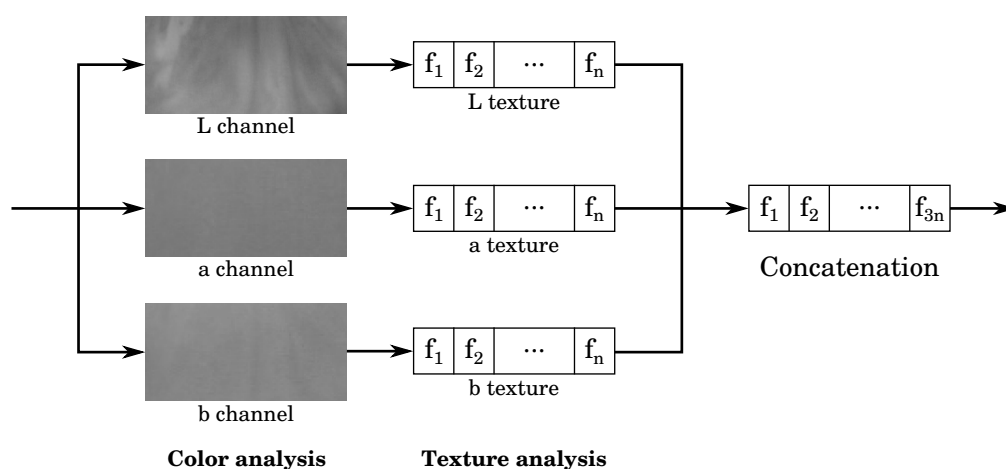


Figure 2.6: Feature vector steps using the Lab color space.

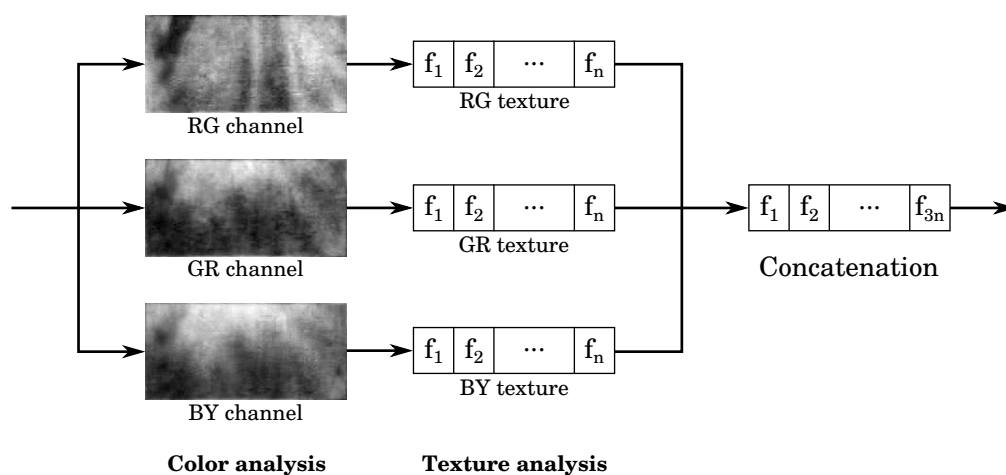


Figure 2.7: Feature vector steps using opponent colors.

dataset. Secondly, all the texture analysis methods were combined for each color space. Next, a support vector machine (Burges, 1998) with radial basis kernel and automatic parameter estimation was trained, using a 10-fold cross validation (see Appendix C). Finally, the effectiveness of the methods were evaluated in terms of the predictive accuracy of the classifier.

Appendix A includes the detailed tables of the results obtained from these experiments. Below, the results obtained with the five texture analysis techniques and the three color spaces, and those obtained by means of the combinations between the texture methods are analyzed.

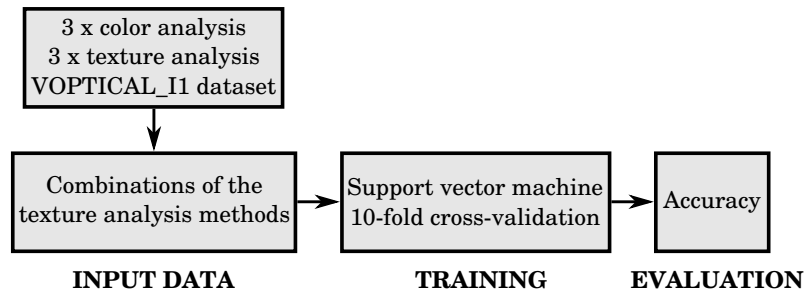


Figure 2.8: Experimental procedure related to the color and texture methods.

Butterworth filters

The first experiment was performed using Butterworth filters and the three color models. A bank of 9 frequency bands filters and histograms composed of 16 bins were considered. Each frequency band was analyzed separately, and the adjacent frequency bands were combined by means of the concatenation of their individual descriptors. This experiment is useful to decide which color space and frequency bands are more appropriate for this task. See Table A.1 in the Appendix A, which shows the results in terms of accuracy for all the frequency band concatenations.

Analyzing these results it can be seen that the intermediate frequencies are more discriminative than the lowest and highest ones; achieving results over a 70% of correct classifications in grayscale. The best combinations provide classification rates higher than 80%. Regarding opponent colors, we can see how color information improves the accuracy of the method compared to grayscale. In this case, the accuracy is almost 90% for the best combinations of frequency bands. Finally, the results show that Lab outperforms opponent colors and produces the best results, which reach classification rate of 93.33%. Table A.1 also shows how the results are quite stable, since there is a wide range of frequency band combinations where the results are over a 90% accuracy. Regarding the number of scales, the best result in grayscale was achieved concatenating the 9 frequencies. In contrast, the other color spaces needed to concatenate only 3 bands to achieve the best results: bands 5 to 7 using Lab, and bands 2 to 4 using opponent colors.

Gabor filters

The second experiment was performed using Gabor filters, and consisted in using a different number of bins to create the uniform histogram which defines the descriptor in grayscale, Lab and opponent colors. Concretely, 3-bin, 5-bin, 7-bin and 9-bin

histograms were analyzed using a bank of 16 Gabor filters. The choice of these histograms is because a greater number of bins does not outperform the results. See Table A.2 in the Appendix A, which shows the results in terms of percentage accuracy for all the histogram sizes.

Analyzing the obtained results it can be seen that all of them achieve around a 90% of correct classifications. In fact, the results are quite stable regardless of the number of bins. The best results have been highlighted for each color space, with maximum accuracy of 95.24% using the Lab color space.

The discrete wavelet transform

The third experiment was performed with the discrete wavelet transform and aimed to analyze not only the behavior of each mother wavelet but also the number of scales. Scales from 1 to 5 were analyzed, and the mother wavelets considered were Haar and Daubechies (Daub4, Daub6 and Daub8). Note that the Haar wavelet is equivalent to Daub2. See Table A.3 in the Appendix A, which shows the results in terms of percentage accuracy for all the scales and mother wavelets.

Analyzing these results it can be seen that the larger scales are more discriminative than the smaller ones. Also, in general terms, the Daub6 wavelet provides the highest accuracy rates. Regarding grayscale, the use of the Haar wavelet achieves results over 89% of correct classifications in several scales. On the other hand, the use of color information improves the results both in Lab and opponent colors. Concerning opponent colors, the accuracy is over 91% using the Daub6 and Daub8 wavelets. However, the best result is obtained with the combination of the Lab color space and Daub6 (94.29%), which is closely followed by Lab and the other wavelets with no significant differences. In addition, the results are quite stable, since there is a wide range of scales for which the accuracy is over 90%, independently of the mother wavelet.

Markov random fields

The fourth experiment was carried out using Markov random fields and aimed at comparing different neighborhoods in the three color spaces. Each distance from 1 to 10 was analyzed individually, as well as the combination of the adjacent distances by means of the concatenation of their descriptors. See Table A.4 in the Appendix A, which shows the results in terms of accuracy for all the distances.

Analyzing the obtained results it can be seen that there is a range of distances between 3 and 6 that achieve over an 80% accuracy. The lowest and highest distances perform worse, which is an indication of the medium size of the texture patterns.

Regarding the use of color information, it does not always outperform grayscale. The best results for each color space have been highlighted, and all of them provide a classification rate of almost 85%. Concretely, the best results in grayscale corresponds to distance 3, to distance 4 in Lab, and to distance 1 in opponent colors which slightly outperform the other color models with an accuracy of 88.57%.

Co-occurrence features

The fifth experiment was related to co-occurrence features, and its target is to analyze the impact of using different distances in the three color spaces. Each distance from 1 to 7 was analyzed individually, as well as the combination of the adjacent distances through the concatenation of their descriptors. See Table A.5 in the Appendix A, which shows the results in terms of accuracy for all the distances.

Analyzing these results it can be seen that the highest distances are more discriminative than the lowest ones, and provide over a 90% of correct classifications in grayscale. The best distance combinations provide classification rates over a 92%. Opponent colors do not outperform grayscale, being the results quite similar. However, these results are improved by the Lab color space. Almost all the distance combinations obtain an accuracy over 90%, and some of them around 95%. The best result in grayscale was obtained using the distance 7, the concatenation of distances 3 to 4 in opponent colors, and the distance 6 in Lab that, once again, is the color space which produced the best results. As well as using Markov random fields, the great behavior of the method using these intermediate distances is an indication of the medium size of the texture patterns.

Combination of texture analysis methods

After analyzing the results obtained with each texture analysis method using the three color spaces previously mentioned, the last experiment was performed in order to check if the categorization accuracy could be increased by combining the texture methods. Concretely, all the possible combinations of the five texture analysis methods were analyzed and, for this task, the best individual results in each color space were considered. In this sense, Table 2.1 shows a summary of the previous five experiments and includes the best result for each pair texture-color. These results are presented in terms of percentage accuracy, and the parameter configuration of each pair is specified in brackets.

Firstly, the texture analysis methods have been combined two by two and the obtained results can be seen in Table A.6. In grayscale, the best individual result corresponds to the co-occurrence features analysis, and is improved by two com-

Table 2.1: Summary of the best combinations for color and texture analysis: SVM classification accuracy (%) and parameter configuration.

	Grayscale	Opp. Colors	Lab
Butterworth filters	83.81 (freqs. 1-9)	90.48 (freqs. 2-4)	93.33 (freqs. 5-7)
Gabor filters	88.57 (3 bins)	88.57 (5 bins)	95.24 (7 bins)
Discrete wavelet transform	89.52 (Haar 3 sc.)	91.43 (Daub6 4 sc.)	94.29 (Daub6 4 sc.)
Markov random fields	83.81 (dist. 4)	84.76 (dist. 1)	83.81 (dist. 3)
Co-occurrence features	92.38 (dist. 7)	92.38 (dists. 3-4)	96.19 (dist. 6)

binations which do not include it. In opponent colors, the best individual result is also obtained using co-occurrence features. However, the combination of Butterworth filters and Markov random fields is the only one which outperforms the co-occurrence features analysis. Regarding the Lab color space, the best individual results is once again provided but co-occurrence features analysis, which combined with the discrete wavelet transform outperforms it.

Secondly, the methods were combined three by three and the results are depicted in Table A.7. Using three method combinations, the best individual results obtained using grayscale are always outperformed. Regarding opponent colors, only one combination (the discrete wavelet transform, Markov random fields and Gabor filters) improves the best individual results provided by co-occurrence features analysis. And finally, all the combinations that outperform the best individual result using the Lab color space include the co-occurrence features which seems to be a key method in the problem at hand.

Next, the methods were combined four by four and the results are presented in Table A.8. These results are quite similar independently of the color space considered, since in all the cases the combinations outperform their respective individual results, and also there are some combinations that improved the best individual result of each color space.

Finally, the five methods were combined and the obtained results are shown in Table A.9. In grayscale and Lab, the individual results are improved by the combination of all methods. However, the combination of all methods using opponents

colors is not able of improving the performance of the best individual results obtained by using the co-occurrence features analysis. On the other hand, a similarity can be found by comparing the three color spaces: the five method combination results are not better than the four method combination results. This fact could be caused by the large dimensionality of the feature vector, which could complicate the classification process.

2.4 Classification

Supervised machine learning is one of the tasks most frequently carried out by so-called *intelligent systems*. Thus, a large number of techniques have been developed based on *artificial intelligence*, such as logic-based algorithms; and *statistics*, such as Bayesian networks (Kotsiantis, 2007). The goal of supervised learning is to construct a classifier than can correctly predict the classes of new samples given training samples of old objects (Mitchell, 1997). The training process consists in learning a mapping between a set of input features and output labels. The resulting classifier is then used to assign class labels to the new instances whose values of the features are known, but the value of the class label is unknown.

2.4.1 Machine learning algorithms

Five popular machine learning algorithms were selected in order to provide different approaches of the learning process (Remeseiro et al., 2012).

Naive Bayes

Naive Bayes (NB) (Jensen, 1996) is an statistical learning algorithm based on the Bayesian theorem and the maximum posteriori hypothesis which can predict class membership probabilities. During the training process, the posteriori probabilities of each class are calculated according to the Bayes' theorem:

$$P(c_j, X) = \frac{P(X, c_j)P(c_j)}{P(X)} \quad (2.21)$$

where c_j is a class and X is a sample. $P(a, b)$ represents the posteriori probability of a conditioned on b , and $P(a)$ represents the priory probability of a .

Given a sample X , the trained classifier will predict that X belongs to the class which has the highest a posteriori probability conditioned on X . That is, X is predicted to belong to the class c_i if and only if:

$$P(c_i, X) > P(c_j, X), j \neq i \quad (2.22)$$

where the class c_i is called the maximum posteriori hypothesis.

This classifier greatly simplify learning by assuming that features are independent of the given class. Although independence is generally a poor assumption, in practice this algorithm competes well with more sophisticated classifiers (N. Friedman, Geiger, & Goldszmidt, 1997). Thus, its main advantage is that it is simple and fast, but its problem lies in it cannot learn interactions between features.

Logistic model tree

Tree induction methods and logistic models are two popular techniques for supervised learning tasks. The combination of these two schemes results in a single tree called logistic model tree (LMT) (Landwehr, Hall, & Frank, 2005), i.e., a tree which contains logistic regression functions at the leaves.

A logistic model tree consists of a tree made up of a set of inner nodes N , and a set of leaves T . Let S denotes the whole instance space, spanned by all attributes present in the data. Then, the tree structure gives a disjoint subdivision of S into S_t regions, and so that every region is represented by a leaf:

$$S = \bigcup_{t \in T} S_t, S_t \cap S_{t'} = \phi \text{ for } t \neq t' \quad (2.23)$$

Unlike ordinary decision trees, the leaves $t \in T$ have an associated logistic regression function f_t instead of a class label. The function f_t takes into account a subset $V_t \subseteq V$ of all attributes, and models the class-membership probabilities as:

$$Pr(G = j | X = x) = \frac{e^{F_j(x)}}{\sum_{k=1}^J e^{F_k(x)}} \quad (2.24)$$

where:

$$F_j = \alpha_0^j + \sum_{k=1}^m \alpha_{v_k}^j \cdot v_k \quad (2.25)$$

Thus, the model represented by the whole logistic model tree is given by:

$$f(x) = \sum_{t \in T} f_t(x) \cdot I(x \in S_t) \quad (2.26)$$

Notice that logistic regression and ordinary decision trees are special cases of logistic model trees: the former is a LMT pruned back to the root, and the latter

is a tree in which $V_t = \phi$ for all $t \in T$. Note also that the main advantage of using logistic regression is that explicit class probability estimations are produced rather than just a classification.

Random tree

Random tree (RT) (Biau, 2012) is a tree randomly constructed from a set of possible trees having K random features at each node. In this context, “at random” means that in the set of trees each tree has an equal chance of being sampled.

In order to construct a random tree, all its nodes are associated with rectangular cells such that at each step of the construction, the collection of cells associated with the leaves forms a partition of $[0, 1]^d$. The root of the tree is $[0, 1]^d$ itself. The following procedure is then repeated $\lceil \log_2 k_n \rceil$, where \log_2 is the base-2 logarithm, $\lceil \cdot \rceil$ is the ceiling function, and $k \geq 2$ a deterministic parameter. The procedure is as follows: at each node, a coordinate of $X = (X^{(1)}, \dots, X^{(d)})$ is selected, with the j -th feature having the probability $p_{nj} \in (0, 1)$ of being selected; next, the split is at the midpoint of the chosen side.

Notice that a randomized tree $r_n(X, \Theta)$, where Θ is a randomizing variable, outputs the average error over all Y_i for which the corresponding vectors X_i fall in the same cell of the random partition as X . Note also that the main advantage of random trees is that they can be generated efficiently.

Random forest

Random forest (RF) (Breiman, 2001) is an effective tool in predictive tasks formed by a combination of tree predictors. Formally, it can be defined as a classifier which consists of a collection of tree-structured classifiers $\{h(x, \Theta_k), k = 1, \dots\}$ where the $\{\Theta_k\}$ are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input x .

Given an ensemble of classifiers $h_1(x), h_2(x), \dots, h_k(x)$, and with the training set randomly drawn from the distribution of the random vector Y, X , the margin function can be defined as follows:

$$mg(X, Y) = av_k I(h_k(X) = Y) - \max_{j \neq Y} av_k I(h_k(X) = j) \quad (2.27)$$

where $I(\cdot)$ is the indicator function. The margin measures the extent to which the average number of votes at X, Y for the right class exceeds the average vote for any other class. Thus, the larger the margin, the more confidence in the classification.

The generalization error is given by:

$$PE^* = P_{X,Y}(mg(X, Y) < 0) \quad (2.28)$$

where the subscripts X, Y indicate that the probability is over the X, Y space.

For a large number of trees, it follows from the *strong law of large numbers* and the tree structure (Breiman, 2001) that as the number of trees increases, for almost surely all sequences Θ_1, \dots PE^* converges to:

$$P_{X,Y}(P_{\Theta}(h(X, \Theta) = Y) - \max_{j \neq Y} P_{\Theta}(h(X, \Theta) = j) < 0) \quad (2.29)$$

where $h_k(X) = h(X, \Theta_k)$ in random forests.

This theorem explains why random forest does not overfit as more trees are added, which is the main advantage of this method. However, it produces a limiting value of the generalization error.

Support vector machine

Support vector machine (SVM) (Burges, 1998) is based on the statistical learning theory and revolves around the notion of a “margin”, either side of a hyperplane that separates two classes. If the training data is linearly separable, then a hyperplane that separates two classes can be defined as:

$$w \cdot x + b = 0 \quad (2.30)$$

where x are the samples, w is the normal to the hyperplane and $\frac{b}{\|w\|}$ is the perpendicular distance from the hyperplane to origin. The aim of SVMs is to orientate this hyperplane in such a way as to be as far as possible from the closest members of both classes, which means selecting the variables w and b so that the training data can be described by:

$$y_i(x_i \cdot w + b) - 1 \geq 0 \quad (2.31)$$

where x_i is the i -th sample, and y_i its class. From all the possible hyperplanes, SVMs try to find the one that maximizes the margin. Vector geometry shows that the margin is equal to $\frac{1}{\|w\|}$, so maximizing it is equivalent to minimizing $\|w\|$.

Most real world problems involve non-separable data for which no hyperplane exists that successfully separates two classes. In this case, the idea is to map the input data onto a higher dimensional space and define a separating hyperplane there. This higher-dimensional space is called the transformed feature space and it is obtained using *kernel functions*.

SVM necessarily reaches a global minimum and avoids ending in a local minimum, which may happen in other algorithms. They avoid problems of overfitting and, with an appropriate kernel, they can work well even if the data is not linearly separable. However, the SVM methods are binary so multi-class problems have to be transformed to a set of multiple binary problems.

2.4.2 Experimental study

The target here is to test the significance of the differences among classifier accuracies, and so five experiments have been performed using the five texture analysis methods, and the five classifiers previously mentioned.

The experimental procedure is detailed in Figure 2.9. Firstly, the three color analysis and the five texture analysis methods were applied to the VOPTICAL_I1 dataset. Secondly, the five classifiers are trained using a 10-fold cross validation (see Appendix C). Note that a SVM with radial basis kernel and automatic parameter estimation was considered according to (Remeseiro et al., 2012). Finally, the effectiveness of the methods were evaluated in terms of the predictive accuracy of the classifier. In addition, a statistical comparison of classifiers was performed based on the Lilliefors test for normality, the ANOVA test and the Tukey's method for multiple comparison (see Appendix D).

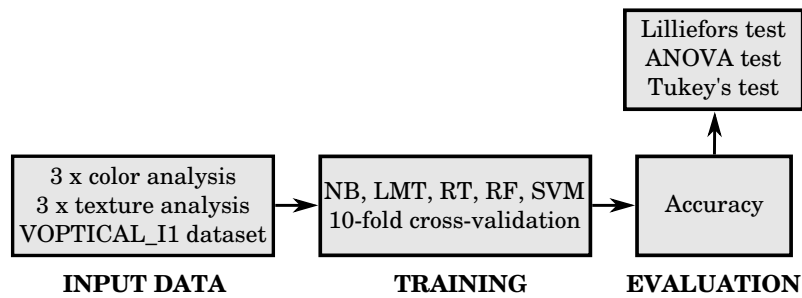


Figure 2.9: Experimental procedure related to the different classifiers.

Appendix A includes the detailed tables of the results for these five experiments. Below, the results obtained with the five texture analysis techniques and the three color spaces using the five machine learning algorithms are analyzed.

Butterworth filters

The first experiment was performed using Butterworth filters, and analyzes each frequency band separately (see Table A.10). The Lilliefors test for normality accepted

the null hypothesis that the data came from a normal distribution in all the color spaces, and so the ANOVA test was performed (see Table A.11). In grayscale, the ANOVA test rejected the null hypothesis and the Tukey's test concluded that there are not significant differences among SVM, LMT and RF. On the other hand, the ANOVA test rejected the null hypothesis using opponent colors, and the Tukey's test concluded that there are significant differences among SVM and all the classifiers but LMT. Finally, the ANOVA test accepted the null hypothesis for the Lab color space, which means that no classifier performs significantly different from the others.

Gabor filters

The second experiment analyzes the Gabor filters using 4 different histogram sizes (see Table A.12). In grayscale, the Lilliefors test accepted the null hypothesis and then, the ANOVA test concluded that there are significant differences among the classifiers (see Table A.13). Concretely, the SVM is significantly different from the others classifiers according to the Tukey's test. In opponent colors, the SVM did not pass the normality test and was not considered in the ANOVA test. This test concluded that there are significant differences among classifiers and the Tukey's test selected the RF and LMT as the statistically better ones. Regarding the Lab color space, the Lilliefors test rejected the null hypothesis for the NB classifier, which was not included in the ANOVA test. On the other hand, the ANOVA test rejected the null hypothesis and the multiple comparison test selected the SVM as the classifier with significant differences with respect to the others.

The discrete wavelet transform

The third experiment aimed to analyze the discrete wavelet transform using 5 scales and the Daub6 as the mother wavelet (see Table A.14). In the three color spaces, the Lilliefors test accepted the null hypothesis while the ANOVA test rejected it (see Table A.15). Regarding the Tukey's method in grayscale, it concludes that there are significant differences among SVM, NB, LMT and RT, but not among SVM and RF; so SVM and RF are the best classifiers in this case. On the other hand, the Tukey's test concluded that the SVM is significantly different from the other four classifiers using the Lab color space and opponent colors.

Markov random fields

The fourth experiment consisted in analyzing the Markov random fields method with 10 different neighborhoods (see Table A.16). In grayscale, the Lilliefors test accepted the null hypothesis and the ANOVA test rejected it (see Table A.17). Finally, the multiple comparison test concluded that the SVM has significant differences with all the classifiers. In Lab and opponent colors, the results obtained with the NB classifier are not normally distributed. The NB classifier produced the poorer results in terms of percentage accuracy so it was eliminated from the experiment. Using the other four classifiers, the ANOVA test rejected the null hypothesis in both color spaces (see Table A.17). Finally, the multiple comparison test concluded that SVM has significant differences with the other classifiers.

Co-occurrence features

The last experiment analyzes the co-occurrence features and considers 7 distances separately (see Table A.18). In the three color spaces, the Lilliefors test accepted the null hypothesis and the ANOVA test rejected the null hypothesis (see Table A.19). The Tukey's test also concluded the same in the three color spaces: there are significant differences among the SVM, which is the method that performs best, and the other four classifiers.

Summary

As a summary, Table 2.2 shows the most competitive classifiers for each texture extraction method in the three color spaces, according to the experiments performed. Analyzing these results, it can be seen that SVM outperforms the other classifiers in most cases, since it fits better the boundaries between classes. Thus, it should be established as the most suitable method for the problem at hand.

2.5 Conclusions

A methodology for tear film assessment has been presented, based on the automatic classification of the Tearscope images into one of the Guillon categories. It locates the region of interest of an input image, analyzes its low-level features through different color spaces and texture analysis methods, and finally classifies it into one of the categories by using machine learning algorithms. The obtained results show how the automatic classification, with the developed strategy, is feasible with results over 80% of accuracy in all the methods tested. This accuracy validate the general

Table 2.2: Summary with the most competitive classifiers using the five texture extraction methods and the three color spaces. (*) SVM produces higher accuracies than LMT and RF, but it could not be included in the experiment because its data did not come from a normal distribution according to the Lilliefors test.

Texture analysis	Color analysis		
	Grayscale	Opponent colors	Lab
Butterworth filters	SVM, LMT, RF	SVM, LMT	-
Gabor filters	SVM	LMT, RF (*)	SVM
The discrete wavelet transform	SVM, RF	SVM	SVM
Markov random fields	SVM	SVM	SVM
Co-occurrence features	SVM	SVM	SVM

strategy, regardless of the color model and texture descriptor used, and even other alternative techniques were considered.

In general terms, the use of color information improves the results compare to grayscale because some lipid layers contain, not only morphological features, but also color features. All the texture analysis methods perform quite well providing results over the 90% in some cases, but co-occurrence features analysis generates the best results. Although Markov random fields use information of the pixel's neighborhood, as the co-occurrence features technique does, this method does not work so well because the statistics proposed by Haralick et al. provide much more information. In short, the combination of co-occurrence features and the Lab color space produces the best classification results with maximum accuracy over 96%.

On the other hand, the texture analysis methods have been combined in order to improve the accuracy. With this combination, grayscale can reach the accuracy obtained with Lab and opponent colors, with an accuracy over 96%. In regard to the texture methods, co-occurrence features analysis provides the best individual results in the three color spaces as stated above. However, this method has been known to be slow and, despite an optimization of the method was implemented based on (Clausi & Jernigan, 1998), it presents an unacceptable extraction time (several tens of seconds). The combinations of methods which do not include the co-occurrence features analysis allows to get about the same accuracy that using only this time-consuming technique, and in less time (under ten seconds). In this manner, the co-occurrences features analysis becomes not essential in the texture analysis step.

Regarding the machine learning algorithms, the SVM produces the best results independently of the texture extraction method and the color space, compared with

other four classifiers. In order to check if the differences among classifiers were significant, an statistical comparison was performed. For this task, the Lilliefors test was applied to assess the normality of the results in terms of percentage accuracy. Based on the conclusions of this test, the ANOVA test was subsequently applied to check whether the differences among classifiers were significant or not. If they were significant, the Tukey's test was applied to decide which classifiers were significantly different from the others. As a result, the SVM classifier presents significant differences compared to the other classifiers and so it is considered as the most competitive method. However, the LMT should be also considered because it is the second most competitive method according to the results obtained, and it has an advantage compared to SVM: it does not need parameter tuning.

In clinical terms, the manual process done by experts can be automated with the benefits of being faster and unaffected by subjective factors. The system is able to provide unbiased results with maximum accuracy over 96%, which relieves optometrists from this tedious task. Several experienced optometrists have performed this task by hand in order to compare their classifications, and analyze their level of agreement (García-Resúa et al., 2013). The agreement between these subjective observers was established in the range from 91% to 100%. Therefore, the clinical significance of the results obtained with the proposed methodology should be highlighted: the 96% of accuracy provided by the system is in the same range that the agreement between experts, which ratifies the correct performance of the system.

Chapter 3

Dimensionality reduction

The complexity of any classification process depends on the number of input attributes, apart from the own complexity of the corresponding classifier. This determines both memory and time complexity, and also the necessary number of training samples to train the classifier. According to the approach presented in the Chapter 2 for the automatic tear film classification, the co-occurrence features technique (Haralick et al., 1973), as a texture extraction method, and the Lab color space (McLaren, 1976) provide the highest discriminative power from a wide range of methods analyzed. However, the best accuracy results are obtained at the expense of a too long processing time because many features have to be computed, which also means too much memory. This fact makes this methodology unfeasible for practical applications and prevents its clinical use. Therefore, different dimensionality reduction techniques are applied in an attempt to decrease the number of features and, consequently, the computational (memory and time) requirements without compromising the classification performance.

This chapter tries to reduce the complexity of the problem through dimensionality reduction techniques, which can be divided into two main groups: feature extraction methods which form fewer, new features from the original attributes; and feature selection methods which choose a subset of relevant features pruning the rest.

3.1 Feature extraction

Feature extraction (Alpaydin, 2010) is a special form of dimensionality reduction, which transforms the data in the high-dimensional space to a space of fewer dimensions. That is, it finds a new set of k input attributes, which are combinations of the original d attributes ($k < d$), and it maintains a high percentage of the original information. The relevant information from the input data is extracted to perform

the desired task using this reduced representation instead of the full size input.

The most popular feature extraction methods are *principal component analysis* (PCA) and *linear discriminant analysis* (LDA), which are both linear projection methods, unsupervised and supervised respectively. In this study, PCA has been chosen to perform some experiments (Remeseiro, Penas, et al., 2013), and so it will be subsequently described.

3.1.1 Principal component analysis

Principal component analysis (PCA) (Jolliffe, 1986) is a feature transformation technique widely used for dimensionality reduction. It is an unsupervised method and so it does not use the output information; the criterion to be maximized is the variance. In fact, it reduces the dimensionality of the input data by performing a variance analysis between factors. According to that, it is useful when there is a large number of variables and there could be some redundancy in those variables. In this case, redundancy means that some of the variables are correlated and it would be possible to reduce the variables into a smaller number of principal components.

In mathematical terms, this procedure uses an orthogonal transformation to convert a set of values of possibly correlated variables into a set of values of uncorrelated variables known as principal components. This transformation is defined in such a way that the first principal component captures the highest possible variance, and each successive component captures the highest remaining variance under the constraint of being orthogonal to all the preceding components.

3.1.2 Experimental study

The objective is to analyze the impact of using *principal component analysis* on the percentage accuracy. Thus, a total of 3 experiments were performed in order to compare the results obtained with and without applying PCA, one per each color analysis approach.

The experimental procedure is detailed in Figure 3.1. Firstly, the three color analysis and the five texture analysis methods were applied to the VOPTICAL_I1 dataset. Secondly, all the texture analysis methods were combined for each color space. Then, the PCA technique is applied to all the combination of methods. As the variance is the criterion to be maximized in this method, different variances values were considered ranging from 90% to 99%. Next, a support vector machine (Burges, 1998) with radial basis kernel and automatic parameter estimation was trained, using a 10-fold cross validation (see Appendix C). Finally, the effectiveness

of the PCA technique was evaluated in terms of the predictive accuracy of the classifier, and the number of extracted features.

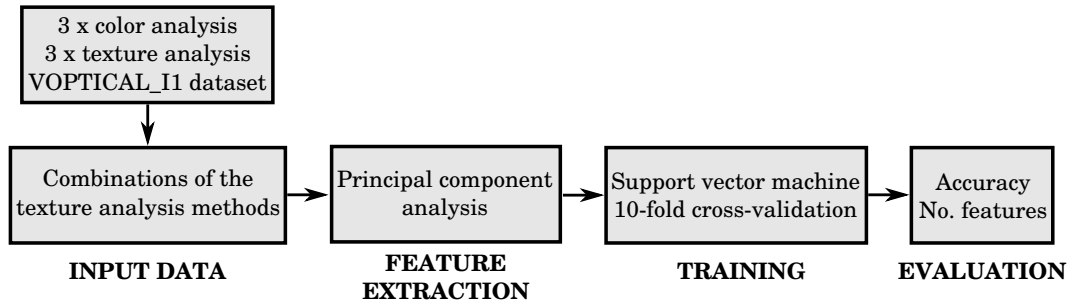


Figure 3.1: Experimental procedure related to the PCA technique.

Appendix A includes the detailed tables of the results for these experiments. Below, the results obtained with and without the use of PCA using the five texture analysis techniques and the three color spaces are analyzed.

The first experiment was performed using grayscale and all the combinations of the texture analysis methods. Table A.20 shows some representative results of this experiment. Analyzing these results, it can be seen that the accuracy keeps around the same percentage despite the great reduction of the number of variables, which reaches the 85% in most cases. Furthermore, this accuracy is maintained in most of cases, and improved in some combinations.

The second experiment was performed using opponent colors and all the combinations of the texture analysis methods. Table A.21 shows some notable results of the experiment. In this case, most of the combinations outperform the classification rates obtained without applying PCA and, in the rest of cases, there is no degradation in performance despite the impressive reduction of the number of variables which surpasses the 95% in several combinations.

The last experiment was performed using the Lab color space and all the combinations of the texture analysis methods. Table A.22 shows the results of some selected combinations. There is no degradation in performance in most combinations, even in some of them the classification rates are improved in spite of the dimensionality reduction which rounds the 90%.

As a conclusion, the use of PCA allows the reduction in memory requirements by transforming the input space and produces no degradation in performance. However, as a transformation is applied, the whole feature vector has to be calculated and so there is no reduction in time.

3.2 Feature selection

Machine learning can take advantage of feature selection to reduce the number of features so as to improve the performance of automatic classifiers (Guyon, Gunn, Nikravesh, & Zadeh, 2006). Feature selection methods can be divided into three main models: filters, wrappers and embedded methods (Guyon et al., 2006). The filter model relies on general characteristics of the data (correlation, entropy, etc.) to evaluate and select feature subsets without involving any learning algorithm or prediction model. On the other hand, wrapper models use a specific prediction method as a black box to score subsets of features as part of the selection process. Finally, embedded methods perform feature selection as part of the training process of the prediction model. By having some interaction with the classifier, wrapper and embedded methods tend to give better performance results than filters, at the expense of a higher computational cost. Also, it is well-known that wrappers have the risk of overfitting when having more features than samples (Loughrey & Cunningham, 2005), as it is the case in this research. Trying to overcome this limitation, some preliminary tests have been performed in this research using a wrapper approach with sequential forward search, however the performance obtained was not good. The poor behavior showed by wrappers in this kind of scenarios, together with the significant computational burden required by this approach, prevent their use in this research. Therefore, filters were chosen because they allow for reducing the dimensionality of the data without compromising the time and memory requirements of machine learning algorithms.

3.2.1 Filters

Among the broad suite of methods present in the literature, three filters were chosen (Bolon-Canedo et al., 2012; Remeseiro, Bolon-Canedo, et al., 2014) and are subsequently presented.

Correlation-based feature selection

Correlation-based feature selection (CFS) is a simple multivariate filter algorithm that ranks feature subsets according to a correlation based heuristic evaluation function (M. A. Hall, 1999). The bias of the evaluation function is toward subsets that contain features that are highly correlated with the class and uncorrelated with each other. On the one hand, irrelevant features should be ignored because they will have low correlation with the class. On the other hand, redundant features should be screened out as they will be highly correlated with one or more of the remaining

features. The acceptance of a feature will depend on the extent to which it predicts classes in areas of the instance space not already predicted by other features. CFS feature subset evaluation function is:

$$M_S = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}}, \quad (3.1)$$

where M_S is the heuristic ‘merit’ of a feature subset S containing k features, $\overline{r_{cf}}$ is the mean feature-class correlation ($f \in S$) and $\overline{r_{ff}}$ is the average feature-feature intercorrelation. The numerator of this equation can be thought of as providing an indication of how predictive of the class a set of features is, whilst the denominator of how much redundancy there is among the features.

Consistency-based filter

The consistency-based filter (Dash & Liu, 2003) evaluates the worth of a subset of features by the level of consistency in the class values when the training instances are projected onto the subset of attributes. The algorithm generates a random subset S from the number of features in every round. If the number of features of S is lower than the best current set (S_{best}), the data with the features prescribed in S is checked against the inconsistency criterion. If its inconsistency rate is below a pre-specified one, S becomes the new S_{best} .

The inconsistency criterion, which is the key to the success of this algorithm, specifies to what extent the dimensionally reduced data can be accepted. If the inconsistency rate of the data described by the selected features is smaller than a pre-specified rate, it means the dimensionally reduced data is acceptable.

INTERACT

The INTERACT algorithm (Zhao & Liu, 2007) is a subset filter based on symmetrical uncertainty (SU) (Press, Flannery, Teukolsky, & Vetterling, 1986), which is defined as the ratio between the information gain (IG) and the entropy (H) of two features, x and y :

$$SU(x, y) = 2 \frac{IG(x/y)}{H(x) + H(y)} \quad (3.2)$$

where the information gain is defined as:

$$IG(x/y) = H(y) + H(x) - H(x, y) \quad (3.3)$$

where $H(x)$ and $H(x, y)$ are the entropy and joint entropy, respectively.

INTERACT also includes the consistency contribution (c-contribution). The c-contribution of a feature is an indicator about how significantly the elimination of that feature will affect consistency. The algorithm consists of two major parts. In the first part, the features are ranked in descending order based on their SU values. In the second part, features are evaluated one by one starting from the end of the ranked feature list. If the c-contribution of a feature is lower than an established threshold, the feature is removed, otherwise it is selected. The authors stated in (Zhao & Liu, 2007) that INTERACT can thus handle feature interaction, and efficiently selects relevant features.

3.2.2 Experimental study

The objective is to find which feature selection filter performs better, i.e. it is the most appropriate for the texture analysis methods considered and the Lab color space, which is the color model which performs best according to previous experiments (see Chapter 2). In this sense, two experiments were carried out: one which evaluates the effectiveness of the filters using three performance measures, and other which is focused on a particular case of study.

The experimental procedure is detailed in Figure 3.2. Firstly, the Lab color space and the five texture analysis methods were applied to the VOPTICAL_I1 and VOPTICAL_R datasets. Secondly, the three feature selection filters were applied to the VOPTICAL_I1 dataset in order to provide the subset of features which properly describe the given problem. Next, a support vector machine (Burges, 1998) with radial basis kernel and automatic parameter estimation was trained, using a 10-fold cross validation (see Appendix C). Finally, the effectiveness of the filters were evaluated in terms of three performance measures (accuracy, robustness and feature computing time).

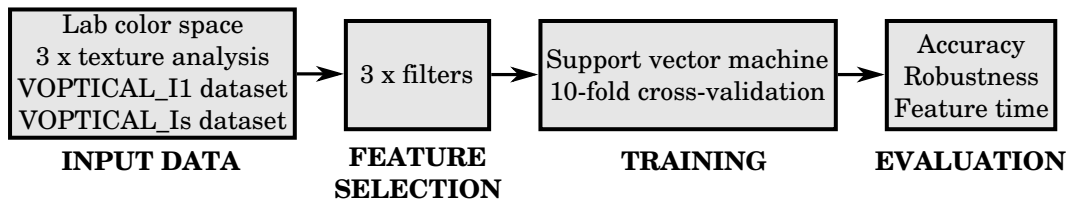


Figure 3.2: Experimental procedure related to the feature selection filters. Experimentation was performed on an Intel®Core™i5 CPU 760 @ 2.80GHz with RAM 4 GB.

These three performance measures are described as follows:

- The *accuracy* is the percentage of correctly classified instances on a dataset with optimum illumination.
- The *robustness* is the classification accuracy in a noisy dataset, i.e. its accuracy when the images in the dataset show illuminations outside the optimum range. This measure is related to the generalization ability of the method when handling noisy inputs. Notice that the higher the robustness, the higher the generalization performance.
- The *feature computing time* is the time that the texture analysis methods take to compute the selected features of a single image. Note that this does not include the training time of the classifier, which is not relevant for practical applications because the classifier will be trained off-line. This also applies to feature selection, which is a pre-processing step performed off-line.

Table 3.1 contains the parameter configurations considered in this step. The combinations of all the individual parameters have been considered whenever it is possible. Thus, the feature selection filters can find the best subset of features from a complete set which includes all the information available.

Table 3.1: Parameter configurations of the texture analysis methods using the Lab color space, and number of features

Texture analysis	Configuration	No. features
Butterworth filters	Frequencies 1-9	432
Gabor filters	7-bin histograms	336
The discrete wavelet transform	Daub6, 5 scales	81
Markov random fields	Distances 1-10	660
Co-occurrence features	Distances 1-7	588

Bear in mind that the column *None* in the tables of this section shows the results when no feature selection was performed. The number of features selected by each of the three feature selection filters is summarized in Table 3.2, which also includes the percentage of the initial features selected in each case. On average, CFS, consistency-based filter (*Cons*) and INTERACT (*INT*) allows the elimination of the 94.4%, 98.1% and 96.2% of the features, respectively.

Table 3.2: Number of features, and percentage of the initial features selected.

Texture analysis	Feature selection filter			
	None	CFS	Cons	INT
Butterworth filters	432	26 (6.02%)	6 (1.39%)	14 (3.24%)
Gabor filters	336	29 (8.63%)	7 (2.08%)	18 (5.36%)
Discrete wavelet transform	81	12 (14.81%)	8 (9.88%)	11 (13.58%)
Markov random fields	660	24 (3.64%)	13 (1.97%)	15 (2.27%)
Co-occurrence features	588	27 (4.59%)	6 (1.02%)	21 (3.57%)

Classification accuracy

Table 3.3 shows the test accuracies for all pairwise texture analysis and feature selection methods after applying the SVM classifier over the VOPTICAL_I1 dataset. The best result for each texture model is marked in bold face. As can be seen, all texture analysis techniques perform quite well providing results over 84% accuracy. Gabor filters and co-occurrence features without feature selection outperform the other methods. Although Markov random fields use information of the pixel's neighborhood, as the co-occurrence features do, this method does not work so well because the statistics proposed by Haralick et al. (Haralick et al., 1973) provide much more textural information. Regarding feature selection, it outperforms primal results in two out of five methods (Butterworth filters and Markov random fields), while accuracy is almost maintained in co-occurrence features analysis when CFS is applied. As conclusions, the best result is obtained by using the co-occurrence features or Gabor filters, when no feature selection is performed (95.24%). Closely, the discrete wavelet transform when feature selection is not applied, and co-occurrence features with CFS (94.29%). Notice that although these results do not mean a degradation in performance despite the reduction of the input space, the goal here is to reduce the processing time whilst maintaining accuracy.

Robustness to noise

Table 3.4 shows the robustness of the five different methods over the VOPTICAL_Is dataset. The co-occurrence features analysis obtains remarkable better results than the remainder methods, and it is the only one which provides values of robustness over 90% for some configurations. In particular, the best result is obtained by using co-occurrence features when CFS filter is used (92.36%). In relative terms, the co-occurrence features method deteriorates its mean classification accuracy by 2.66%

Table 3.3: Mean test classification accuracy (%).

Texture analysis	Feature selection filter			
	None	CFS	Cons	INT
Butterworth filters	91.42	93.33	83.81	86.67
Gabor filters	95.24	91.43	86.67	86.67
Discrete wavelet transform	94.29	91.43	89.52	80.95
Markov random fields	84.76	85.71	83.81	75.24
Co-occurrence features	95.24	94.29	86.67	93.33

(mean difference between the values contained in Tables 3.3 and 3.4). However, the remainder methods deteriorate their mean classification accuracy by between 6.89% and 8.23%. Note also that the illumination levels affect the robustness in different degrees. The brighter the illumination, the lower the robustness to noise. This also happens to practitioners when performing this task by hand. For this reason, their experience to control the illumination level during the acquisition stage is cornerstone for ensuring good classification performance.

Table 3.4: Robustness: mean test accuracy (%) in the noisy dataset.

Texture analysis	Feature selection filter			
	None	CFS	Cons	INT
Butterworth filters	88.18	84.98	71.92	79.56
Gabor filters	89.90	85.22	69.46	82.51
Discrete wavelet transform	88.92	79.31	79.80	77.34
Markov random fields	83.99	76.35	70.94	70.69
Co-occurrence features	92.17	92.36	85.22	89.16

Feature computing time

Tear film lipid layer classification is a real-time task so the time a method takes to process an image cannot be a bottleneck. After applying feature selection and so reducing the number of input attributes, the time needed for analyzing a single image with any of the five methods was also reduced as can be seen in Table 3.5. In general terms, Butterworth filters, the discrete wavelet transform and Gabor filters take a negligible lapse of time to obtain the features of an image (regardless of whether or

not feature selection is applied as preprocessing step). Moreover, Markov random fields takes a time which could be acceptable for practical applications, even when no feature selection is applied, although it could not work in real time. The co-occurrence features technique has been known to be slow and, despite the authors implemented an optimization of the method based on (Clausi & Jernigan, 1998), it presents an unacceptable computing time. Co-occurrence features analysis is only acceptable for practical applications when consistency-based or INTERACT filters are used. Consistency-based filter selects fewer features (see Table 3.2) and consequently the processing time when this filter is used is smaller.

Table 3.5: Feature computing time in seconds.

Texture analysis	Feature selection filter			
	None	CFS	Cons	INT
Butterworth filters	0.22	0.15	0.04	0.07
Gabor filters	0.42	0.18	0.06	0.11
Discrete wavelet transform	0.18	0.09	0.07	0.08
Markov random fields	13.83	0.50	0.27	0.31
Co-occurrence features	102.18	27.01	0.05	9.86

Co-occurrence features with CFS: a case of study

When using feature selection, features are selected according to some specific criteria depending on the method. Specifically, filters remove features based on redundancy and relevance. However, no one of the methods takes into account costs for computing those features. Note that the cost of obtaining a feature depends on the procedures required to extract it. In this manner, each feature has an associated cost which can be economic, related to a physical risk or computational demanding. This is the case of co-occurrence features, in which the cost of computing the 588 features is not homogeneous. Features are vectorized in groups of 28 related to distances and components in the color space according to Table 3.6. Each group of 28 features corresponds with the mean and range of 14 statistics across the gray level co-occurrence matrices (see Section 2.3.2).

Notice that, when using CFS, the number of features were reduced by 95.41% (from 588 to 27) but the processing time was not reduced proportionally, and is now 27.01s instead of 102.28s (a reduction of 73.57%). This fact clearly shows that extracting some of the 588 features takes longer than others. Some experimentation

Table 3.6: Features within distances and components.

Distance	Component in the color space		
	L	a	b
1	1–28	29–56	57–84
2	85–112	113–140	141–168
3	169–196	197–224	225–252
4	253–280	281–308	309–336
5	337–364	365–392	393–420
6	421–448	449–476	477–504
7	505–532	533–560	561–588

was performed on the time that co-occurrence features analysis takes to calculate each of the 14 statistics. Results disclosed that computing the 14th statistic uses around 96% of the total time. So the key for reducing the processing time is to reduce the number of 14th statistics in the selection.

The 27 features selected by CFS are depicted in Table 3.7, grouped by distance and component in the color space. Four of these features correspond with the 14th statistic and are remarked in bold face. In co-occurrence features, the cost of computing the statistics also depends on the distance and component in the color space. On the one hand, the longer the distance, the larger the number of matrices to compute, and so the higher the processing time. On the other hand, as explained before, in the Lab color space, L represents the luminance while a and b represent the colorimetric components. However, the differences of color have little contrast so the colorimetric components of the Lab color space are minimal. As a consequence, the matrices within components a and b have smaller dimension than the matrices within component L . As expected, the smaller the dimension, the shorter the time to compute a statistic.

Computing the 14th statistics involved in Table 3.7 takes: 4.79s (feature 182), 7.98s (feature 350), 9.59s (feature 434), and 4.58s (feature 546). As can be seen, avoiding computing some of them will entail saving a significantly amount of time (up to 26.94 seconds out of a total time of 27.01 seconds, i.e. the 99.74%). The aim here is to explore the impact of removing some of the 14th statistics selected by CFS in terms of accuracy and time. There are 4 features within the 14th statistic, and so 2^4 different configurations are to be explored. Thus an empirical evaluation of brute force is acceptable. Table 3.8 shows the performance of the different configurations in terms of accuracy and time. Each configuration corresponds with

Table 3.7: Set of the 27 features within distances and components using CFS, in which features corresponding with 14th statistic are marked in bold.

Distance	Component in the color space		
	L	a	b
1	6	50, 54	66
2	91	113, 121, 133	–
3	182	–	230, 237
4	254, 261, 262, 267, 268, 275, 276	–	–
5	350, 359	–	–
6	434	–	492, 502
7	530	546, 553	576

those features selected by CFS removing some 14th statistics. For example, row $CFS-\{182\}$ corresponds with all the features selected by CFS except feature 182.

In terms of accuracy, the best result is achieved in 7 cases, obtaining a 97.14%, which is the highest precision so far. Among these, the best trade-off is attained by $CFS-\{182, 434, 546\}$, employing 8.05 seconds. However, this time is still intractable in a real-time application. Finally, when using CFS without the 14th statistics ($CFS-\{182, 350, 434, 546\}$), the performance in terms of accuracy is slightly decreased with respect to the best result (corresponding to misclassify one sample) but with a very acceptable time (less than 1 second). It is also noticeable that when this approach is compared with CFS (see first row in Table 3.8), the accuracy is improved whilst the time is reduced by 99.74% (from 27.01 to 0.07s).

CFS selects the features based on the correlation with the class and, although redundant features should be screened out, this fact may not happen. In addition, the feature selection filters in general, and CFS in particular, are independent of the classifier. Thus, the predictive accuracy may be different for the same subset of features depending on the classifier considered. And this is exactly what happens when removing the 14th statistic from the subset of selected features.

Even when the cost of the 14th statistic along the distances and components is significant, the effectiveness of CFS filter for selecting the most appropriate features is also remarkable. Further experimentation showed this fact: if only the 14th statistics are removed from the 588 features, the accuracy is 94.29%, i.e. the accuracy is worse than $CFS-\{182, 350, 434, 546\}$. As expected, the time is also worse (0.24s) because of the need of computing more features.

Table 3.8: Performance measures for co-occurrence features with CFS when some of the 14th statistics are excluded from the subset of selected features.

Features	Accuracy(%)	Time(s)
CFS	94.29	27.01
CFS-{182}	97.14	22.22
CFS-{350}	97.14	19.03
CFS-{434}	97.14	17.42
CFS-{546}	95.24	22.43
CFS-{182,350}	97.14	14.24
CFS-{182,434}	97.14	12.63
CFS-{182,546}	96.19	17.64
CFS-{350,434}	96.19	9.44
CFS-{350,546}	95.24	14.45
CFS-{434,546}	95.24	12.84
CFS-{182,350,434}	95.24	4.65
CFS-{182,350,546}	97.14	9.66
CFS-{182,434,546}	97.14	8.05
CFS-{350,434,546}	96.19	4.86
CFS-{182,350,434,546}	96.19	0.07

3.3 Cost-based feature selection

New feature selection methods are continuously emerging, being successfully applied to different areas (Forman, 2003; Inza, Larrañaga, Blanco, & Cerrolaza, 2004). However, the great majority of them only focus on removing unnecessary features from the point of view of maintaining the performance, but do not take into account the possible different costs for computing the features.

Although features with a related cost can be found in many real-life applications, this has not been the focus of much attention for machine learning researchers. To the best knowledge of the authors, there are only a few attempts in the literature to deal with this issue (Feddema, Lee, & Mitchell, 1991; Huang & Wang, 2006; Sivagaminathan & Ramakrishnan, 2007; Min, Hu, & Zhu, 2013). Most of these methods have the disadvantage of being computationally expensive by having interaction with the classifier, which prevents their use in large datasets. A quick examination of the most popular machine learning and data mining tools revealed

that no cost aware methods can be found. Weka (M. Hall et al., 2009) only include some methods which address the problem of cost associated to the instances, not to the features. RapidMiner (Mierswa, Wurst, Klinkenberg, Scholz, & Euler, 2006) includes some methods to handle cost related to features, but they are quite simple.

Consequently, a modification of the well-known filter ReliefF was proposed in (Bolon-Canedo, Remeseiro, Sánchez-Marroño, & Alonso-Betanzos, 2014). This filter was chosen since: it can be applied in many different situations, it has low bias, it includes interaction among features, and it has linear dependency on the number of features. Therefore, the proposed mC-ReliefF will be suitable even for application to datasets with a great number of input features.

3.3.1 mC-ReliefF

Relief (Kira & Rendell, 1992) and its multiclass extension, ReliefF (Kononenko, 1994), are supervised feature weighting algorithms included in the filter approach. The key point is to estimate the quality of attributes according to how well their values distinguish between instances which are near to each other. Therefore, given a randomly selected instance R_i , the Relief algorithm searches for its two nearest neighbors: one for the same class, *nearest hit* H , and the other from the different class, *nearest miss* M .

The ReliefF algorithm is not limited to two class problems, is more robust, and can deal with incomplete and noisy data. As the original Relief algorithm, ReliefF randomly selects an instance R_i , but then searches for k of its nearest neighbors from the same class, nearest hits H_j , and also k nearest neighbors from each one of the different classes, nearest misses $M_j(C)$. It updates the quality estimation $W[A]$ for all attributes A depending on their values for R_i , hits H_j and misses $M_j(C)$. If instances R_i and H_j have different values of the attribute A , then this attribute separates instances of the same class, which clearly is not desirable, and thus the quality estimation $W[A]$ has to be decreased. On the contrary, if instances R_i and M_j have different values of the attribute A for a class then the attribute A separates two instances with different class values which is desirable so the quality estimation $W[A]$ is increased. Since ReliefF considers multiclass problems, the contribution of all the hits and all the misses is averaged. Besides, the contribution for each class of the misses is weighted with the prior probability of that class $P(C)$ estimated from the training set. The whole process is repeated m times, where m is a user-defined parameter, and can be seen in Algorithm 3.1.

Algorithm 3.1: Pseudo-code of ReliefF algorithm

Data: training set D , iterations m , attributes a

Result: the vector W of estimations of the qualities of attributes

```

1 set all weights  $W[A] := 0$ 
2 for  $i \leftarrow 1$  to  $m$  do
3   randomly select an instance  $R_i$ 
4   find  $k$  nearest hits  $H_j$ 
5   for each class  $c \neq class(R_i)$  do
6     from class  $c$  find  $k$  nearest misses  $M_j(c)$ 
   end
  end
7 for  $f \leftarrow 1$  to  $a$  do
8    $W[f] :=$ 
    $W[f] - \frac{\sum_{j=1}^k diff(f, R_i, H_j)}{(m \cdot k)} + \frac{\sum_{c \neq class(R_i)} \left[ \frac{P(c)}{1 - P(class(R_i))} \sum_{j=1}^k diff(f, R_i, M_j(c)) \right]}{(m \cdot k)}$ 
  end

```

The function $diff(A, I_1, I_2)$ calculates the difference between the values of the attribute A for two instances, I_1 and I_2 . For nominal attributes, it is defined as:

$$diff(A, I_1, I_2) = \begin{cases} 0; & value(A, I_1) = value(A, I_2) \\ 1; & otherwise \end{cases}$$

The modification of ReliefF here proposed, mC-ReliefF, consists in adding a term to the quality estimation $W[f]$ to take into account the cost of the features:

$$W[f] := W[f] - \frac{\sum_{j=1}^k diff(f, R_i, H_j)}{(m \cdot k)} + \frac{\sum_{c \neq class(R_i)} \left[\frac{P(c)}{1 - P(class(R_i))} \sum_{j=1}^k diff(f, R_i, M_j(c)) \right]}{(m \cdot k)} - \lambda \cdot C_f, \quad (3.4)$$

where C_f is the cost of the feature f , and λ is a free parameter introduced to weight the influence of the cost in the quality estimation of the attributes. When $\lambda > 0$, the greater the λ the greater the influence of the cost.

3.3.2 Experimental study

The aim of the experiment is to study the behavior of the proposed mC-ReliefF under the influence of the λ parameter. It is expected that the larger the λ , the

lower the cost and the higher the error, since increasing λ gives more weight to cost at the expense of reducing the importance of the relevance of the features. The statistical analysis performed could help the user to choose the value of λ .

The experimental procedure is detailed in Figure 3.3. Firstly, the Lab color space and the co-occurrence features analysis were applied to the VOPTICAL_I1. Secondly, the proposed mC-ReliefF was applied over the dataset using different values of the λ parameter. Next, a support vector machine (Burges, 1998) with radial basis kernel and automatic parameter estimation was trained, using a 10-fold cross validation (see Appendix C). Finally, the effectiveness of the method was evaluated by calculating the total cost of the selected features and the classification error. In addition, a Kruskal-Wallis statistical test and a Tukey's test were run for multiple comparison (Hsu, 1996) on the cost and errors obtained.

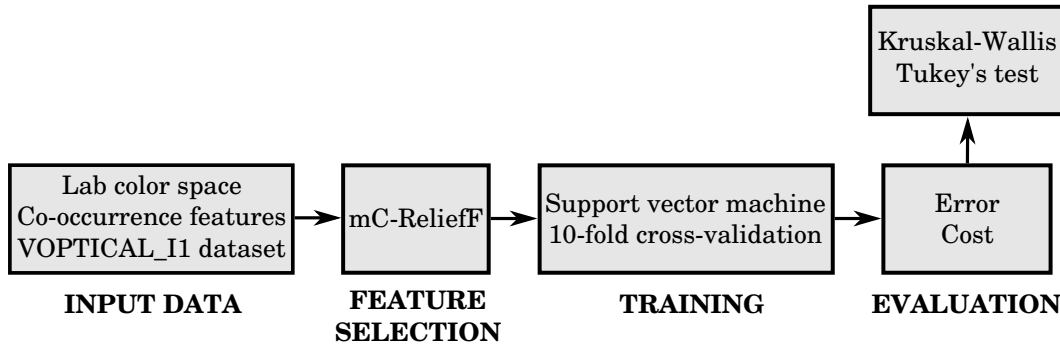


Figure 3.3: Experimental procedure related to the mC-ReliefF algorithm. Experimentation was performed on an Intel®Core™i5 CPU 760 @ 2.80GHz with RAM 20 GB.

The adequacy of mC-ReliefF is tested on the tear film lipid layer classification using the co-occurrence features analysis and the Lab color space, since their combination produces the best performance in terms of accuracy (see Chapter 2). Distances from 1 to 7 in the co-occurrence features method and the 3 components of the Lab color space are considered, so the size of the final descriptor obtained from an input image is: 28 features \times 7 distances \times 3 components = 588 features. Features are vectorized in groups of 28 related to distances and components in the color space. In addition, each group of 28 features corresponds with the mean and range of the 14 statistical measures calculated across the gray level co-occurrence matrices. The cost of computing these features is not homogeneous, since it was shown that computing the so-called 14th statistic takes a great percentage of the total time. Therefore, the dataset considered has a very variable cost (in this case, computational time) associated to the input features.

Figure 3.4 (left) shows the average error and cost after performing a 10-fold cross-validation for VOPTICAL_I1 dataset for different values of λ , for three different sets of features. As expected, when λ increases, the cost decreases and the error either raises or is maintained. Regarding the different subsets of features, the larger the number of features, the higher the cost. The Kruskal-Wallis statistical test run on the results demonstrated that there are no significant differences among the errors achieved using different values of λ , whilst using a $\lambda > 0$ decreases significantly the cost. This situation happens when retaining 25, 35 and 50 features.

Trying to shed light on the issue of which value of λ is better for the problem at hand, the Pareto front (Teich, 2001) for each alternative is showed in Figure 3.4 (right). In multi-objective optimization, the Pareto front is defined as the border between the region of feasible points, for which all constraints are satisfied, and the region of infeasible points. In this case, solutions are constrained to minimize classification error and cost. In Figure 3.4 (right), points (values of λ) in the Pareto front are marked with a red circle. All those points are equally satisfying the constraints, and it is decision of the user if he/she prefers to minimize either the cost or the classification error. On the other hand, choosing a value of λ outside the Pareto front would imply to chose a worse solution than any in the Pareto front.

Table 3.9 reports the classification error and cost (in the form of time) for all the Pareto front points. Notice that as a 10-fold cross-validation was performed, the final subset of selected features is the union of the features selected in each fold, and that is why the number of features in column 5 differs from the one in the first column. Even so, the reduction in the number of features is considerable.

As expected, the higher the λ , the higher the error and the lower the time. The best result in terms of classification error was obtained with $\lambda = 0$ when retaining 50 features per fold. In turn, the lowest time was obtained with $\lambda = 30$ when retaining 25 features per fold, but at the expense of increasing the error in 8.54%. In this situation, the most reasonable decision would be to choose a trade-off between cost and error. The error obtained with $\lambda = 1$ when retaining 35 features is 7.55%, which is slightly higher than the best one but no significant differences were found between them. With this combination the time required is 306.53 milliseconds, which although is not the lowest time, it is still under 1 second. The time required by previous approaches which deal with tear film lipid layer classification prevented their clinical use because they could not work in real time, since computing the whole set of features from a single image took 38 seconds. Thus, since this is a real-time scenario in which reducing the computing time is a crucial issue, having a processing time under 1 second leads to a significant improvement. In this manner,

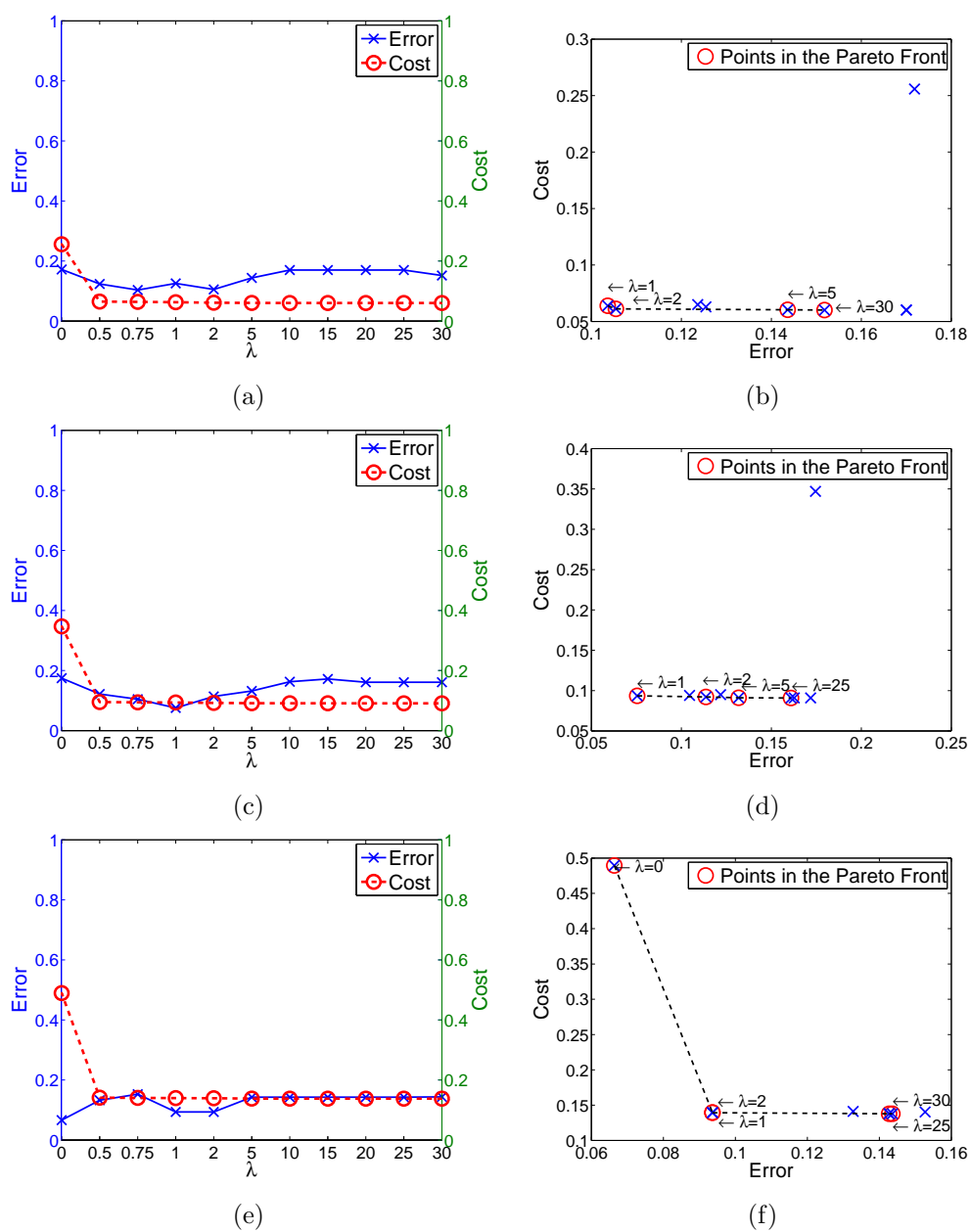


Figure 3.4: From left to right: error / cost plots and Pareto front of the VOPTICAL_I1 dataset for different values of λ , and different number of selected features (25, 35 and 50)

Table 3.9: Mean classification error(%), time (milliseconds), and number of features in the union of the 10 folds for the Pareto front points.

Feats	λ	Error	Time	Feats union
25	0.75	10.36	208.68	30
	2	10.55	206.46	30
	5	14.36	197.22	29
	30	15.18	174.35	26
35	1	7.55	306.53	43
	2	11.36	328.24	46
	5	13.18	273.11	39
	25	16.09	249.92	36
50	0	6.64	1377.04	82
	1	9.36	397.70	55
	2	9.36	412.14	57
	25	14.27	364.45	51
	30	14.36	364.45	51

the methodology for tear film lipid layer classification could be used in the clinical routine as a support tool to diagnose dry eye syndrome.

3.4 Conclusions

A new step of dimensionality reduction was included in the methodology to automatically classify tear film lipid layer patterns. Feature extraction and feature selection methods were applied in order to deal with this step. The PCA technique, a feature extraction method, was applied as a first approach to the problem. The use of this technique allows to reduce dimensionality of the feature vectors up to the 90%, which reduces the memory requirements without impacting the accuracy. In fact, the best result obtained (99.05% of accuracy) corresponds to the combination of the co-occurrence features analysis, the discrete wavelet transform, and the Lab color space after applying the PCA technique with a variance of the 96%.

Although the memory requirements have been reduced using the PCA technique, the time required still prevents their clinical use. To solve this problem, feature selection techniques are applied and so, when an input is decided to be unnecessary, the time used in order to compute it can be saved. Thus, it plays a crucial role

since it reduces the number of input features and also the processing time. Three of the most popular feature selection filters were considered: CFS, consistency-based and INTERACT. They were tested on five popular texture analysis methods and the Lab color space. Results obtained with this new step surpass previous results in terms of processing time whilst maintaining accuracy.

Finally, a modification of the ReliefF filter for cost-based feature selection, called mC-ReliefF, was proposed. ReliefF is a well-known filter, which has proven to be effective in diverse scenarios and includes interaction among features. The extension proposed consists of allowing ReliefF to solve problems where it is interesting not only to minimize the classification error, but also to reduce costs associated to input features. For this purpose, a new term is added to the function which updates the weights of the features so as to be able to reach a trade-off between the relevance of a feature and the cost that it implies. The mC-ReliefF method was applied aiming at reducing the time required to automatically classify the tear film lipid layer patterns. In this scenario the time required to compute the features prevented clinical use because it was too long to allow the software tool to work in real time. The proposed method permits to significantly decrease the required time in over 90%, from 38 seconds to less than 1 second, while maintaining the accuracy.

As a summary, it should be highlighted that the *ad-hoc* feature selection process based on the CFS filter, which reduces the number of features from 588 to 23 with no degradation in performance, is the one that produces the best balance between accuracy and processing time. Concretely, it allows the automation of the manual process with maximum accuracy over 97% and processing time under 1 second. Thus, it is completely recommended the use of the proposed methodology for clinical purposes as a supporting tool to diagnose dry eye syndrome.

Appendix E presents a systematic study of a complete set of machine learning techniques applied to tear film classification, and provides a detailed ranking of configurations. Note that the wide set of techniques used in this study define a 96-alternative configurations in total. Decision-making methods and a conflict handling procedure were used to obtain this ranking list of alternatives based on a total of 7 performance measures, such as accuracy, precision or training time. However, this study does not include the time needed for computing the feature vector, which seems to be key in the use of the system in clinical routines. In addition, the *ad-hoc* solution does not appear in this study as only general techniques were included. Anyway, two of the most similar alternatives to the *ad-hoc* solution are in the first quartile (positions 13th and 22nd), and so it can be concluded that the proposed solution is absolutely valid and competitive.

Chapter 4

Tear film distribution maps

The spatial heterogeneity of the tear film lipid layer (see Figure 4.1) makes the classification of a patient's image into a single Guillon category, as previous approaches do, not always possible. In this manner, the classifications provided by the previous approaches could be little reliable. Alternatively, performing local analysis of the images in order to detect multiple categories per patient would be more accurate. Furthermore, this kind of analysis would be useful to discern different local states, and thus different tear film distribution maps.

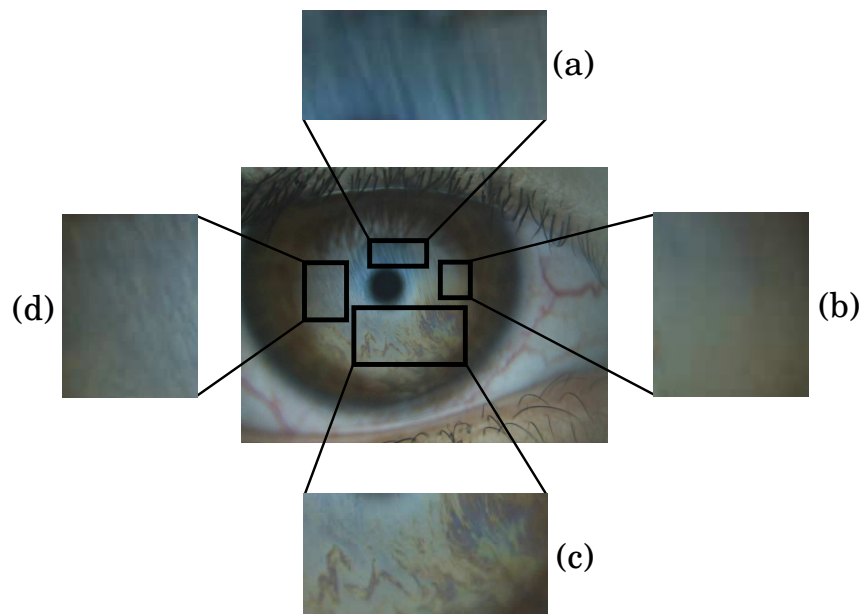


Figure 4.1: Example of the heterogeneity of a patient's tear film lipid layer: (a) wave, (b) amorphous, (c) color fringe, and (d) closed meshwork.

In computational terms, creating a tear film map involves a high increase in the memory and time requirements. This increase is because the previous approach for global tear film classification has to be applied at a local level, and so the feature vectors of hundred of windows have to be calculated. Notice that this problem is manageable thanks to the optimization proposed in Chapter 3, and based on dimensionality reduction techniques. In addition, the increase in computational requirements is compensated by the information obtained with the tear film maps, since this alternative way of analyzing the images provides a detailed distribution of the interference patterns over the tear film lipid layer.

Next sections presents the general methodology to create tear film maps, step by step. In addition, three different alternatives for the main step of the proposed methodology are proposed.

4.1 Optimal window size

Texture segmentation, and in this case the creation of tear film distribution maps, consists in splitting an image into regions of uniform texture. This task is usually performed by applying two stages: the features which characterize each texture are computed, and the obtained features are used to determine uniform regions that allow the segmentation of the image. However, the quality of the final result greatly depends on the size of the regions, i.e. windows, that are analyzed by both stages. On the one hand, it is desirable to use large window sizes since they contain more information than the small ones, and so it is possible to obtain a good texture characterization. On the other hand, finding precise localizations of boundary edges between adjacent regions is a fundamental goal for the segmentation task, and can only be ensured with relatively small windows. Therefore, a certain trade-off regarding window size must be made.

The features are obtained through the approach proposed for tear film classification (see Figure 2.1), which is now applied over local windows instead over the whole region of interest (see Figure 4.2). Consequently, the optimal window size has to be determined, i.e., the minimum window size which allows a precise segmentation and maintains the texture well-defined (Remeseiro, Ramos, Barreira, Mosquera, & Yebra-Pimentel, 2013).

4.1.1 Experimental study

The goal is to determine the optimal window size, and so an experiment was carried out using different sizes and analyzing their impact in the classification accuracy.

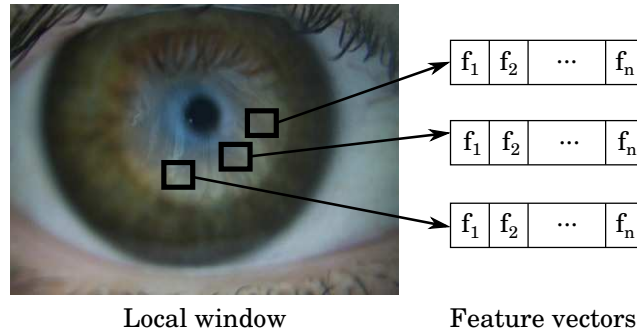


Figure 4.2: Local windows of an input image, and their feature vectors obtained using the previous approach.

The experimental procedure is detailed in Figure 4.3. Firstly, a set of square windows with sizes from 64 to 16 pixels are extracted from the VOPTICAL_R dataset. Note that only the areas in which the three optometrists marked the same category were considered in this stage. Next, a support vector machine (Burges, 1998) with radial basis kernel and automatic parameter estimation was trained, using a 10-fold cross validation (see Appendix C). Finally, the results are evaluated in terms of the percentage accuracy of the classifier.

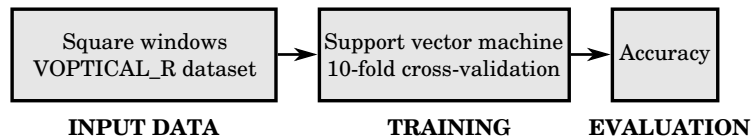


Figure 4.3: Experimental procedure related to the optimum window size.

Figure 4.4 represents the relation between the predictive accuracy of the classifier and the window size. As can be seen, the accuracy for the bigger windows remains almost stable but for the smaller ones, the smaller the window the lower the accuracy. According to these results, the window size selected for image segmentation was 32×32 pixels.

4.2 Research methodology

The proposed methodology is composed of five main steps (see Figure 4.5). From an input image acquired with the Tearscope Plus, some low-level features are obtained from its small windows located at the ROI and so a great amount of information is obtained in order to create a tear film map.

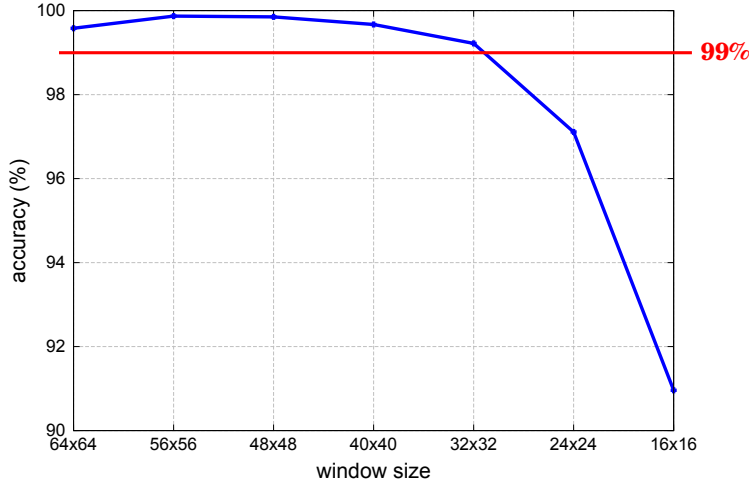


Figure 4.4: Window size vs. accuracy.

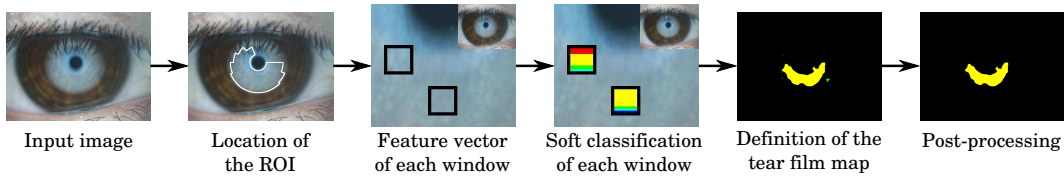


Figure 4.5: Steps of the research methodology to create tear film distribution maps.

These five steps will be subsequently presented in depth, including the experimentation performed. Roughly speaking, the steps are as follows:

1. *Location of the region of interest.* This step aims at finding the area of the input image which corresponds to the whole tear film. This area is known as the region of interest, where the following analysis will take place.
2. *Feature vector.* The low-level features of the region of interest are extracted in this stage based on color and texture information, the two discriminant features of the Guillon categories.
3. *Soft classification.* For each feature vector, its class-membership probabilities are computed using a support vector machine.
4. *Definition of the tear film map.* A tear film map is created in this step, and represented by a labeled image which illustrates the spatial distribution of the lipid layer patterns. These labels correspond to one of the patterns, or represents the background of the image.

5. *Post-processing.* This last step is performed to eliminate the small regions which may appear in the tear film maps.

4.2.1 Location of the region of interest

Input images acquired with the Tearscope Plus include irrelevant areas for tear film segmentation, such as the sclera or the eyelids. Previous approaches located the *region of interest* (ROI) as a rectangle in the bottom part of the iris. Nonetheless, in this case the analysis is taken over the whole tear film and so a new process to locate the ROI is presented (Remeseiro, Mosquera, Penedo, & García-Resúa, 2014).

The whole tear film can be perceived with the best contrast in the green channel of the input image in RGB, so only this single channel will be considered in this stage. First, the green channel is thresholded using its histogram. Then, the *normalized cross-correlation* (Russ, 1999) is applied to the thresholded image, using circles as templates which cover the different pupil sizes. Thus, the circle with the maximum cross-correlation value allows to locate the pupil of the image. Next, a new circle with the same center than the previous one and a radius n times larger is created in order to delimit the area around the pupil. This new circle is used as a first approach to the ROI (see Figure 4.6).

On the other hand, the tear film area is lighter than the iris and the pupil which surround it. In this way, a second approach to the ROI can be determined by finding those pixels whose gray level is greater than a threshold $th = \mu - p \times \sigma$, where μ is the mean value of the gray levels of the image, σ is its standard deviation and p is a weight factor empirically determined.

Since some images can include irrelevant regions, such as eyelashes or shadows cast by them, the morphological operator of *erosion* (Gonzalez & Woods, 2008) is applied in order to eliminate them from this second approach to the ROI (see Figure 4.6). Finally, the logical *AND* operator between the two approaches is calculated. This region is likely to be free of irrelevant features and so, in most cases, could be the final ROI. Despite that, the length of the eyelashes in some cases and specially the irregular shape of this ROI motivate a final adjustment: the biggest circle concentric to the pupil is “divided” in sixteen quadrants and, for each one, the minimum radius is considered in order to simplify the final ROI.

4.2.2 Feature vector

Once the ROI is located, the windows with a specified size inside it are analyzed and a descriptor per window is obtained. This descriptor is a quantitative vector

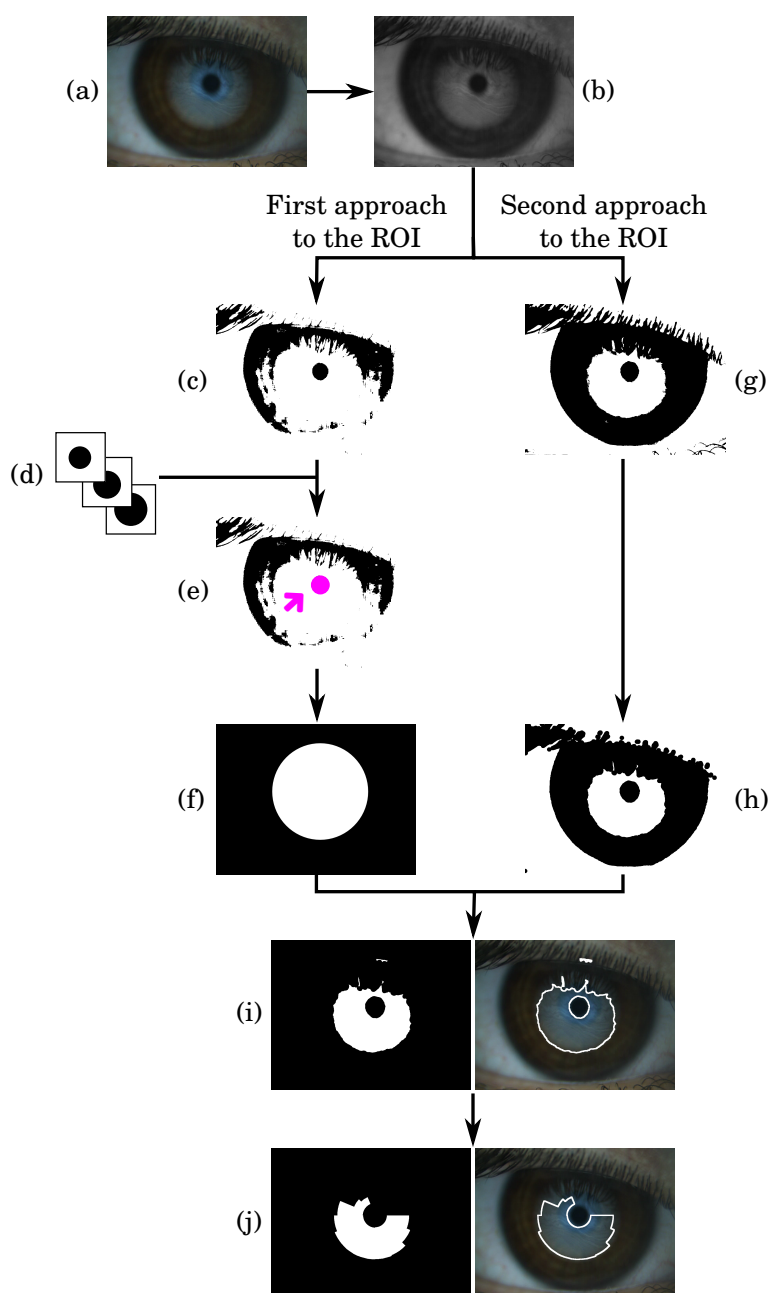


Figure 4.6: Steps for the location of the ROI. (a) Input image acquired with the Tearscope Plus. (b) Green channel of the input image. (c) Thresholded image using the histogram of the green channel. (d) Set of circle-shaped templates. (e) Location of the pupil using the normalized cross-correlation. (f) Resized circle obtained from the pupil approach. (g) Thresholded image using the mean and the standard deviation. (h) Thresholded image after eroding it. (i) Preliminary ROI after applying the AND operator, and its location over the input image. (j) Final ROI after the adjustment, and its location over the input image.

composed of 23 features proposed in (Bolon-Canedo et al., 2012) (see Chapter 3), and obtained as follows:

1. Color analysis. The *Lab* color space (McLaren, 1976) is used to obtain color information, since its use is appropriate in combination with texture analysis.
2. Texture analysis. The *co-occurrence features* technique (Haralick et al., 1973) is used to obtain texture information, since it is the most appropriate method for the problem at hand.
3. Feature selection. The *correlation-based feature selection* (CFS) (M. A. Hall, 1999) was used for feature selection in order to reduce the number of features and, thus, the computational (memory and time) requirements. An *ad-hoc* feature selection process based on this filter was used for dimensionality reduction, so the descriptor with color texture features was reduced, from 588 to 23 features, with no degradation in performance.

4.2.3 Soft classification

From each window located at the ROI, a descriptor is obtained and a *support vector machine* (SVM) (Burges, 1998) is used to compute its class-membership probabilities. Note that partial class memberships are used in soft classification to model uncertain labeling and mixtures of classes. An SVM is used as the machine learning algorithm based on previous results (Remeseiro et al., 2012) (see Chapter 2).

4.2.4 Definition of the tear film map

In this step, a tear film map is obtained using three different approaches: a decision voting system, a weighted voting system and a seeded region growing algorithm. These three approaches will be subsequently explained. Broadly speaking, a tear film map is a labeled image which represents the spatial distribution of the lipid layer patterns. Each label corresponds to one of the Guillon categories or represents the background of the image. In this manner, each tear film map contains several colors which indicate the category of each pixel: red means open meshwork, yellow means closed meshwork, green means wave, cyan means amorphous, blue means color fringe and black means background.

Decision voting system

A first attempt to segment tear film images into the Guillon categories was proposed in (Remeseiro, Ramos, et al., 2013), in order to analyze the feasibility of the problem.

In this preliminary approach, the soft classification is really a hard classification since the maximum probability provided by the classifier is considered. Therefore, the output of the classification process is just a single category.

Once the descriptor of each window located in the ROI is calculated, its category is predicted using a *support vector machine* (SVM) (Burges, 1998) previously trained. In this manner, each pixel in the window receives a vote for the predicted category. As windows are overlapped, each pixel belongs to several windows and so has several votes, which can correspond to different categories. Thus, a decision voting system is necessary in order to obtain the segmented image: for each pixel of the ROI, the number of votes for each category is counted and the pixel is assigned to the most voted category.

Due to the hard classification, in addition to the Guillon categories, this approach includes the “background” as a complementary category. This category is needed since there are no class-membership probabilities, and thus it is necessary to represent those areas of the image where the system does not detect any pattern.

The Algorithm 4.1 shows the whole process of creating a tear film map using the proposed voting system.

Algorithm 4.1: Pseudo-code of the decision voting system.

Data: input image I , minimum perimeter m

Result: output image O (its labels $\in [0, n]$ indicate the classes, where 0 is the background)

```

1  $ROI := locate\_roi(I)$ 
2 initialize matrix of votes  $V := 0$ 
3 for each window  $w \in ROI$  do
4    $feats := compute\_features(w)$ 
5    $CP := classify(feats)$ 
6    $i := index(max(CP))$ 
7   for each pixel  $p \in w$  do
8      $V[p][i] ++$ 
9   end
10 end
11 initialize output image  $O := 0$ 
12 for each pixel  $p \in ROI$  do
13    $i := index(max(V[p]))$ 
14    $O[p] := i$ 
15 end

```

Weighted voting system

The main problem with the previous approach is the necessity of using the so-called *background* category, in addition to the five categories defined by Guillon. This category represents the areas of an image in which no interference pattern has been identified. Since the samples of this unreal category have a high level of variability, they cannot be defined by uniform texture, color features and so the accuracy of a classifier may be affected. Furthermore, it uses the class predicted by the machine learning algorithm to segment the images, regardless of the probability of the prediction. Thus, the proposed method will provide segmented images by means of a weighted voting system which takes into account the multiclass probabilities, and a minimum threshold to confirm the identification of the Guillon categories. This fact makes unnecessary the use of the unreal background category.

The weighted voting system was proposed for tear film segmentation (Remeseiro, Mosquera, et al., 2014), although it could be adapted to any image segmentation problem where the classes can be represented by a set of features and classified by a soft classifier. In the problem at hand, the method considers the class-membership probabilities of each window in the ROI, and every pixel in this window receives a vote associated to each class c :

$$v_c = \omega_1 \cdot p_c + \frac{\omega_2 \cdot p_c}{d} \quad (4.1)$$

where p_c is the probability to belong to the class c , d is the distance from the pixel to the center of the window, and ω_1 and ω_2 weight the probability and the distance, respectively. The idea is the vote depends not only on the probability of belonging to the corresponding class, but also on the distance to the center of the window since in this area the pattern is better defined than in the boundaries of the window.

On the other hand, the maximum vote that every pixel in this window can receive, assuming maximum probability, is also calculated:

$$v_{max} = \omega_1 + \frac{\omega_2}{d} \quad (4.2)$$

All the windows in the ROI are considered in this algorithm and, therefore, windows are overlapped. For this reason, each pixel belongs to several windows, and so the votes received from each category are added up. Thus, each pixel will have a set of final votes corresponding to each class and its maximum final votes. First, only the final votes of the classes are considered in order to select the most voted class. Then, the pixel is assigned to this class only if its final number of votes is higher than the maximum number of votes weighted by a threshold th . Note that

this threshold is used to distinguish the real classes from the background.

The Algorithm 4.2 shows the whole process to create a tear film distribution map using the proposed weighted voting system.

Algorithm 4.2: Pseudo-code of the weighted voting system.

Data: input image I , number of classes n , weights ω_1 and ω_2 , threshold th , minimum perimeter m

Result: output image O (its labels $\in [0, n]$ indicate the classes, where 0 is the background)

```

1   $ROI := locate\_roi(I)$ 
2  initialize matrix of votes  $V := 0$ 
3  initialize vector of maximum votes  $V_{max} := 0$ 
4  for each window  $w \in ROI$  do
5       $feats := compute\_features(w)$ 
6       $CP := classify(feats)$ 
7      for each pixel  $p \in w$  do
8           $d := distance(p, center(w))$ 
9          for  $k \leftarrow 1$  to  $n$  do
10              $v = \omega_1 \cdot CP[k] + \frac{\omega_2 \cdot CP[k]}{d}$ 
11              $V[p][k] += v$ 
12             end
13              $v_{max} = \omega_1 + \frac{\omega_2}{d}$ 
14              $V_{max}[p] += v_{max}$ 
15         end
16     end
17 initialize output image  $O := 0$ 
18 for each pixel  $p \in ROI$  do
19      $v := max(V[p])$ 
20      $i := index(max(V[p]))$ 
21     if  $(v \geq th \cdot V_{max}[p])$  then
22          $O[p] := i$ 
23     end
24 end

```

Seeded region growing

Seeded region growing performs a segmentation of an image with respect to a set of initial points, known as seeds. Given the seeds, which can be manually or automatically selected, the algorithm finds a tessellation of the image into regions. The idea is to analyze each connected component of seeds, through an iterative process, and perform the growing only if the components satisfy a homogeneity criterion.

The original method was presented in (Adams & Bischof, 1994), as applied to grayscale images. An adapted version of this classic algorithm was proposed in (Remeseiro, Mosquera, & Penedo, n.d.) as applied to images based on the class-membership probabilities provided by a soft classifier. The objective is to create tear film distribution maps which represent the spatial distribution of the lipid layer patterns. The description of the new proposal is divided in two parts: the automatic search of the seeds over an input image, and the region growing from the seeds.

The Algorithm 4.3 shows the automatic search of seeds. It consists in analyzing the windows of the ROI in order to calculate their feature vectors, and their corresponding class-membership probabilities (see lines 3 and 4). Then, the maximum class-membership probability is calculated and compared with the seed threshold α . If the probability is greater than the threshold, then the center of the window becomes a seed and so is added to the list of seeds L (see lines from 5 to 12).

Once the seeds are calculated, the process of growing is carried out to get the final regions, as can be seen in Algorithm 4.4. Firstly, the pixels corresponding to the seeds are labeled in the matrix of regions R (see lines from 1 to 5). Then, all the neighbors of the seeds are added to a sorted list SSL (see lines from 6 to 15). This list is sorted based on the homogeneity criterion, which represents the difference between the average class-membership probability of an existing region and the probability of the new pixel which is being analyzed. Thus, the first element in the list will be the one with the minimum δ value, which is defined as:

$$\delta = |CP[i] - mean[i]| \quad (4.3)$$

where $CP[i]$ is the probability of the new element belongs to the class i , and $mean[i]$ is the average probability of belonging to the class i calculated over the pixels which are already labeled as i .

Following, the sorted list SSL is processed until it does not contain any element. Thus, the process subsequently described is applied for each element of the list. The first element is removed from the list, and its neighbors are analyzed (see lines from 17 to 19). If all the neighbors of this element which are already label have the same label, other than the neighbor label, then its δ value previously calculated is

Algorithm 4.3: Pseudo-code of the seed search.

Data: region of interest ROI , number of classes n , seed threshold α

Result: output list of seeds L

```

1 initialize list of seeds  $L := \phi$ 
2 for each window  $w \in ROI$  do
3    $feats := compute\_features(w)$ 
4    $CP := classify(feats)$ 
5    $max := 0$ 
6   for  $k \leftarrow 1$  to  $n$  do
7     if  $CP[k] > max$  then
8        $max := CP[k]$ 
9        $i := k$ 
10    end
11  end
12 if  $max \geq \alpha$  then
13    $seed := create\_seed(w, i)$ 
14    $add(L, seed, i)$ 
15 end
16 end

```

obtained and compared with the β threshold. If δ is lower than the threshold, then the element is labeled with the same label than its neighbors, the average probability of the region is updated, and all the neighbors of the element are added to the SSL list (see lines from 20 to 32). On the other hand, if the neighbors already labeled do not have the same label, then the element is labeled as a boundary (see line 33).

Finally, the tear film map is created by processing the matrix of regions in such a way that those elements which have a label different from the boundary label, are labeled in the output image or tear film map (see lines from 34 to 37).

Once the seeded region growing is performed, the regions may have small holes due to the growing process. In order to homogenize the regions, each hole are "filled" in such a way that its pixels will belong to the region which encloses them.

4.2.5 Post-processing

Once the tear film map is created, small regions may appear in it, which can correspond to false positives or noisy areas. Thus, a post-processing step is performed in order to eliminate them: the regions whose perimeter is less than a minimum

Algorithm 4.4: Pseudo-code of the region growing.**Data:** region of interest ROI , list of seeds L , growing threshold β **Result:** output image O (its labels $\in [0, n]$ indicate the classes, where 0 is the background)

```

1 initialize matrix of regions  $R := 0$ 
2 for each seed  $s \in L$  do
3    $i := getLabel(s)$ 
4    $y := getPos(s)$ 
5    $R[y] := i$ 
6 end
7 initialize sequentially sorted list  $SSL := \phi$ 
8 for each seed  $s \in L$  do
9    $i := getLabel(s)$ 
10   $N = getNeighbors(s)$ 
11  for each neighbor  $n \in N$  do
12     $w := getWindow(n)$ 
13     $feats := compute\_features(w)$ 
14     $CP := classify(feats)$ 
15     $\delta = |CP[i] - mean[i]|$ 
16     $add(SSL, n, i, \delta)$ 
17  end
18 end
19 while notEmpty(SSL) do
20    $y := pushFirst(SSL)$ 
21    $N = getLabeledNeighbors(y)$ 
22    $removeBoundaryNeighbors(N)$ 
23   if sameLabel(N) then
24      $i := getLabel(N)$ 
25      $\delta := getDelta(y)$ 
26     if  $\delta < \beta$  then
27        $R[y] := i$ 
28        $update(mean[i])$ 
29        $N = getNoLabeledNeighbors(y)$ 
30       for each neighbor  $n \in N$  do
31          $w := getWindow(n)$ 
32          $feats := compute\_features(w)$ 
33          $CP := classify(feats)$ 
34          $\delta = |CP[i] - mean[i]|$ 
35          $add(SSL, n, i, \delta)$ 
36       end
37     end
38   else
39      $R[y] := -1$ 
40   end
41 end
42 initialize output image  $O := 0$ 
43 for each pixel  $p \in ROI$  do
44   if ( $R[p] > 0$ ) then
45      $O[p] := R[p]$ 
46   end
47 end

```

perimeter m previously established are eliminated. Notice that this threshold was empirically set to 110 pixels based on the minimum perimeter size of the regions marked by the optometrists in the VOPTICAL_R dataset. See Figure 4.7 as an example of this stage.

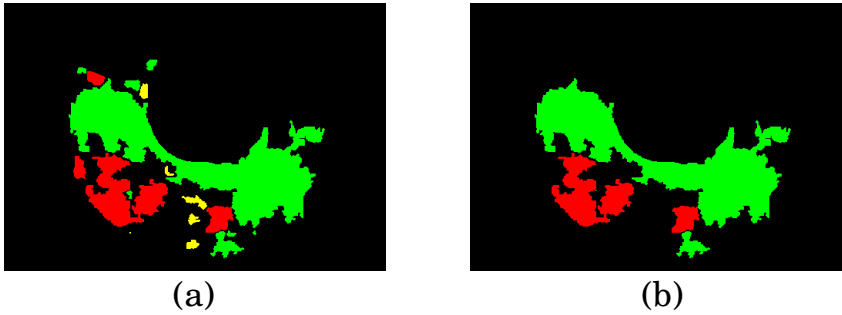


Figure 4.7: (a) Tear film map before the post-processing step. (b) Tear film map after eliminating the small regions.

4.2.6 Experimental study

The objective is to analyze the results obtained with the proposed methodology, and compare the tear film maps obtained with the annotations done by the optometrists. To this end, two different experiments were performed. The first one is related to the decision voting system, a first approach which considers a total six categories (i.e. five real categories defined by Guillon and the background category) and uses hard classification. The second one is related to both weighted voting system and seeded region growing, which consider the five real categories and soft classification.

Experiment 1

The experimental procedure is detailed in Figure 4.8. Firstly, a SVM with radial basis kernel and automatic parameter estimation was trained using representative samples of the six categories considered (including the artificial background category). Notice that only the areas marked by the three optometrists were considered as the five Guillon categories; and regarding the background category, only the areas in which none of the experts marked any category were considered. Secondly, the ROIs of the VOPTICAL_R dataset are located and the feature vectors of their windows are calculated. Next, the classes associated to these quantitative vectors were predicted using the SVM previously trained. Then, the decision voting system was applied and the regions whose contour has a perimeter less than a threshold m

were eliminated. Since this method is a first approach to check the feasibility of the problem, only visual comparisons were done.

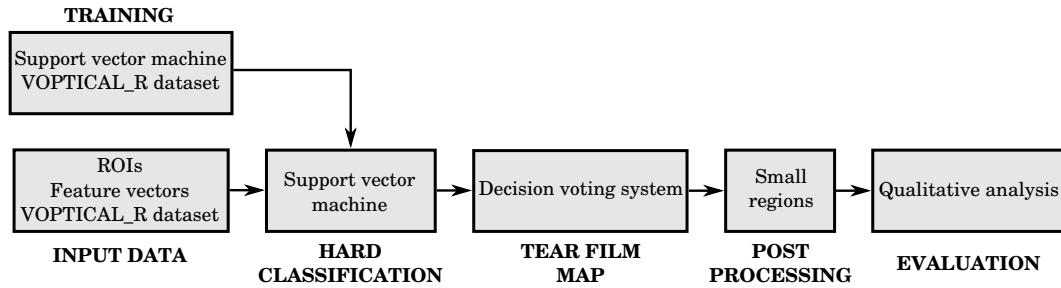


Figure 4.8: Experimental procedure related to the decision voting system.

Figure 4.9 shows the qualitative analysis using five different examples from the VOPTICAL_R dataset, each of them corresponds to a Guillon category. As can be seen, the proposed methodology produces reliable results in comparison with the annotations done by the experts. There are regions of the image in which the experts agree with the Guillon category, whereas there are other regions in which the agreement is non-existent. The same situation happens if the output map is compared with the experts' annotations. Although the windows of the dataset only correspond to areas marked by the three optometrists, the methodology is able to generalize its behavior and can detect other areas marked by just one or two of them.

Experiment 2

The experimental procedure is detailed in Figure 4.10. Firstly, a SVM with radial basis kernel and automatic parameter estimation was trained using representative samples of the five categories considered. Note that, for this task, the samples correspond to areas in which the three optometrists marked the same category. Secondly, the ROIs of the VOPTICAL_R dataset are located and the feature vectors of their windows are calculated. Next, the class-membership probabilities of these quantitative vectors were calculated using the SVM previously trained. Then, both approaches for creating tear film maps were applied: the weighted voting system was applied using different configurations of parameters (ω_1 , ω_2 and th); and the seeded region growing algorithm was applied using different values of the β parameter. Also, the regions whose contour has a perimeter less than a threshold m were eliminated. Finally, the effectiveness of the proposed method was evaluated in terms of the similarity between the system and the three experts considered, not only qualitative but also quantitative. In addition, the process time of the methods

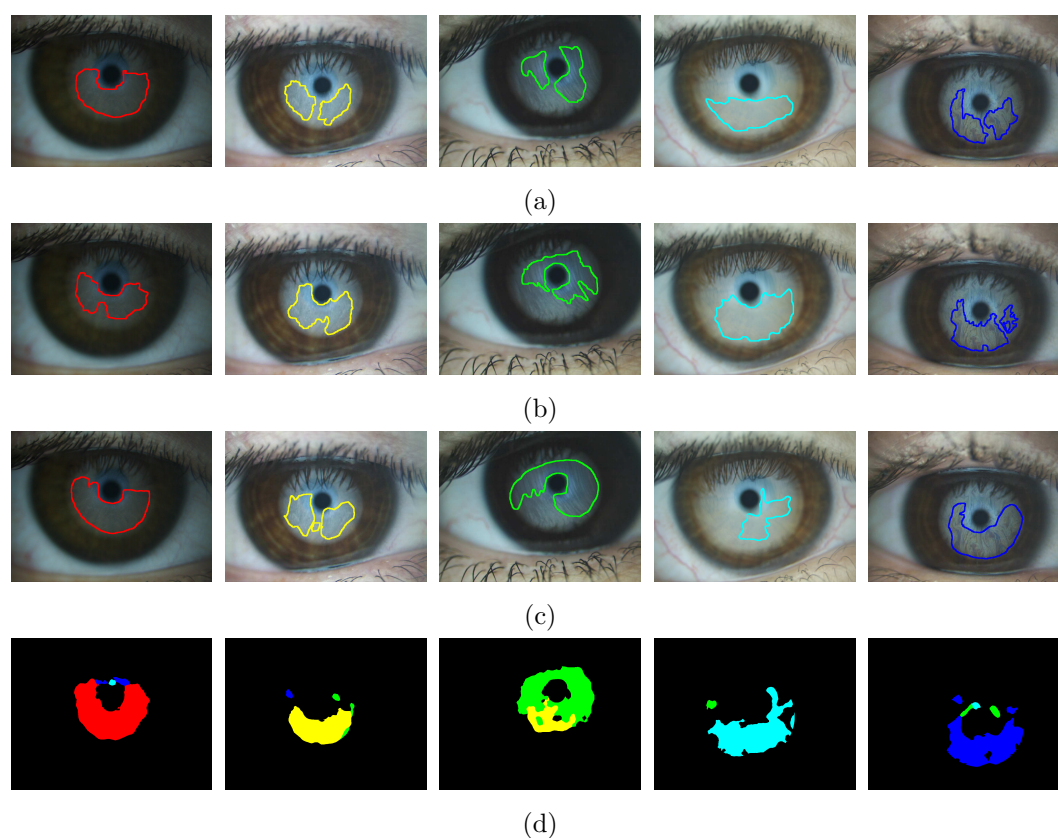


Figure 4.9: Representative images of the VOPTICAL_R dataset. Annotations done by the three optometrists: (a) expert1, (b) expert2, and (c) expert3. (d) Tear film distribution maps obtained with the decision voting system. Note that the relation between colors and categories is: red - open meshwork, yellow - closed meshwork, green - wave, cyan - amorphous, blue - color fringe, and black - background.

was also considered in the validation step. Note that this time does not include the preprocessing step time neither the post-processing step time since both processes are common and independent of the method chosen for creating the tear film map. Furthermore, these two processing times are negligible in comparison with the key procedure of segmentation.

Firstly, the results provided by the proposed methodology will be visually compared with the annotations made by three experienced optometrists. This qualitative comparison is depicted in Figure 4.11, which includes the tear film distribution maps corresponding to five representative images of the VOPTICAL_R dataset, and obtained using both weighted voting system and seeded region growing algorithms. If the regions marked by the experts are compared, it can be seen that they agree

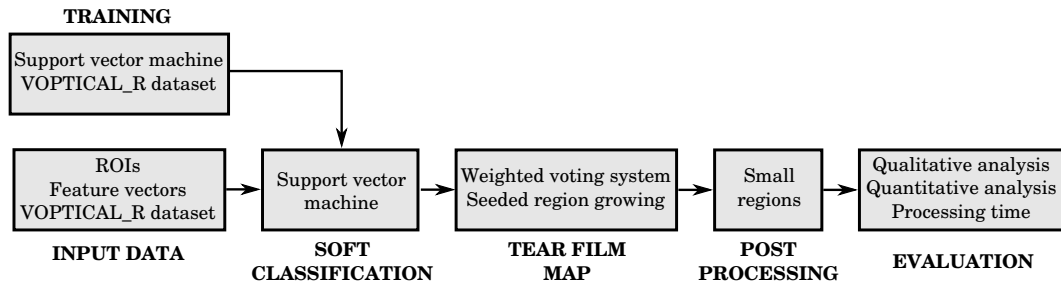


Figure 4.10: Experimental procedure related to the weighted voting system and the seeded region growing algorithm.

in some areas but they disagree in other ones. And the same fact can be appreciated if the tear film maps are analyzed, since some of their regions match with the optometrists' areas and others do not.

The results obtained with these two methods were compared not only graphically, but also in a quantitative way. For this task, the regions marked by the system were compared with the annotations done by the three optometrists, pixel by pixel, and some performance measures were calculated. Before analyzing this comparison, it should be highlighted the difficulty that the optometrists have marking the regions by hand, and the level of disagreement between them. Figure 4.12 illustrates this agreement/disagreement and was obtained by analyzing all the optometrists' annotations of the VOPTICAL_R dataset. For each Guillon pattern, all the pixels marked by the three optometrists in all the images were added up, and the same for those marked by two optometrists or by just only one of them. The graphic shows these values normalized by the total number of pixels per category, and so represents the percentage of pixels associated to each case. This graphic depicts not only the level of agreement, but also the difficulty of the problem. It represents, for each Guillon category, the probability of, given a random pixel classified in this category for a random expert, the other two optometrists or just one of them have been classified this pixel in the same category. As can be seen, the optometrists find more difficult to categorize the color fringe pattern, since the three of them only agree in about a 20% of the pixels marked. In contrast, they fully agree in more than the 50% of the pixels associated to the amorphous pattern. The level of agreement in the other three patterns is in the middle, ranging from 30% and 40%.

Regarding the quantitative analysis carried out, it consists in comparing the results provided by the system with the annotations made by the optometrists. In order to illustrate this comparison, stacked histograms were used to represent the percentage of pixels that the system agrees or disagrees with the optometrists. The

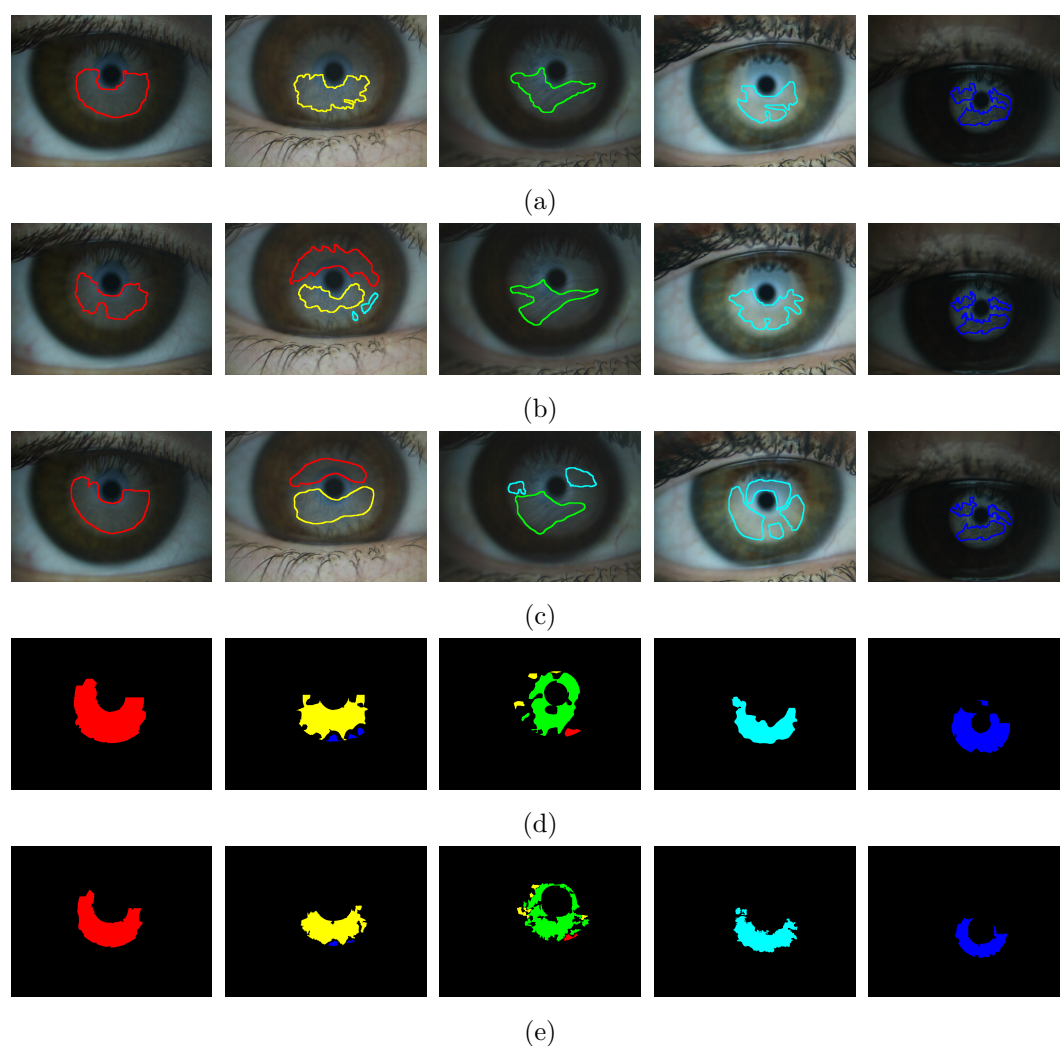


Figure 4.11: Representative images of the VOPTICAL_R dataset. Annotations done by the three optometrists: (a) expert1, (b) expert2, and (c) expert3. (d) Tear film distribution maps obtained with the weighted voting system. (e) Tear film distribution maps obtained with the seeded region growing. Note that the relation between colors and categories is: red - open meshwork, yellow - closed meshwork, green - wave, cyan - amorphous, blue - color fringe, and black - background.

target is not to obtain a particular value of agreement between the system and the experts, but to assess if the system has a behavior equivalent to the behavior of an expert. The comparison consists in analyzing the pixels classified by the system, and checking if they were classified by the experts in the same category. Thus, there will be four different levels of agreement corresponding to those pixels marked by 0, 1, 2 and 3 experts. The agreement with 0 experts means that only the system

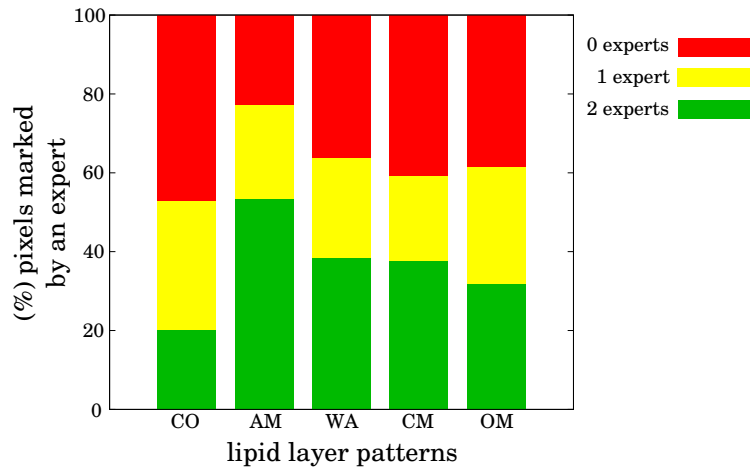


Figure 4.12: Reference graphic which represents the probability of, given a random pixel classified in a given category for a random expert, the other two optometrists (green) or just one of them (yellow) have been classified this pixel in the same category.

marked this area, whilst the agreement with 3 experts means a total agreement between the system and the three experts considered. In addition, the pixels which were not classified by the system were also analyzed and compared with the experts' annotations in a similar way.

In order to evaluate the performance of the system, several measures were calculated from the stacked histograms for each Guillon category. Some basic concepts are explained before defining these measures. The terms true positive (TP), true negative (TN), false positive (FP) and false negative (FN) compare the category predicted by the system with the actual category. True and false refer to if the prediction corresponds to the expectation, while positive and negative refer to the prediction. These basic concepts have to be adapted to the problem at hand to calculate the performance measures. In this sense, positive and negative refer to if the system predicts a Guillon category or the background, respectively. Regarding true and false, the concepts are clear using 3 or 0 experts but the problem lies in the intermediate levels of agreement, which correspond with 1 or 2 experts. Taking into account the difficulty of the problem illustrated in Figure 4.12, it seems reasonable that the agreement with 2 experts is equivalent to agreeing with 3 experts, whilst the agreement with 1 expert is equivalent to agreeing with 0 experts. Thus, the pixels marked by the system and 2 or 3 optometrists will be considered *true positives*, whilst the pixels marked by the system and 0 or 1 expert will be considered *false positives*. In addition, the pixels unmarked by the system and classified into

one of the Guillon categories by 0 or 1 experts will be considered *true negatives*, whilst those classified by 2 or 3 experts will be considered *false negatives*. Using these terms, some performance measures were calculated:

- The *accuracy* is the proportion of true results, both true positives and negatives, i.e. the percentage of correctly classified instances:

$$Acc = \frac{TN + TP}{TP + FP + FN + TN} \quad (4.4)$$

- The *true positive rate*, also called sensitivity or recall, measures the proportion of positives which are correctly classified:

$$TPR = \frac{TP}{TP + FN} \quad (4.5)$$

- The *true negative rate*, also called specificity, measures the proportion of negatives which are correctly classified:

$$TNR = \frac{TN}{TN + FP} \quad (4.6)$$

- The *precision* measures the proportion of the true positives against all the positive results:

$$Prec = \frac{TP}{TP + FP} \quad (4.7)$$

Experiment 2: weighted voting system

Using the weighted voting system, the impact of the different parameter configurations was analyzed. The range of values taken into account for the three parameters is $[0, 1]$. It should be highlighted that at least one of the weights, ω_1 or ω_2 , has to have a non-zero value (see Equation 4.1). Also, if the threshold $th = 1$, the output image contains no information since all the pixels are classified as background. The reason is that a so high threshold implies that most of the class-membership probabilities have the maximum value. Among all the combinations of the three parameters, three of them have been selected to show the validation of the proposed methodology: (i) $[\omega_1 = 1, \omega_2 = 0]$, which only considers the first part of the Equation 4.1, (ii) $[\omega_1 = 0, \omega_2 = 1]$, which only considers the second part of the Equation 4.1, and (iii) $[\omega_1 = 1, \omega_2 = 1]$, which equally considers both parts of the Equation 4.1. In these three cases, the threshold value is $th = 0.9$. Figure 4.13 illustrates the stacked histograms associated to these configurations. If the three configurations are compared, no significant differences can be appreciated which means that considering or not one of the terms of the Equation 4.1 is not too relevant. The reason

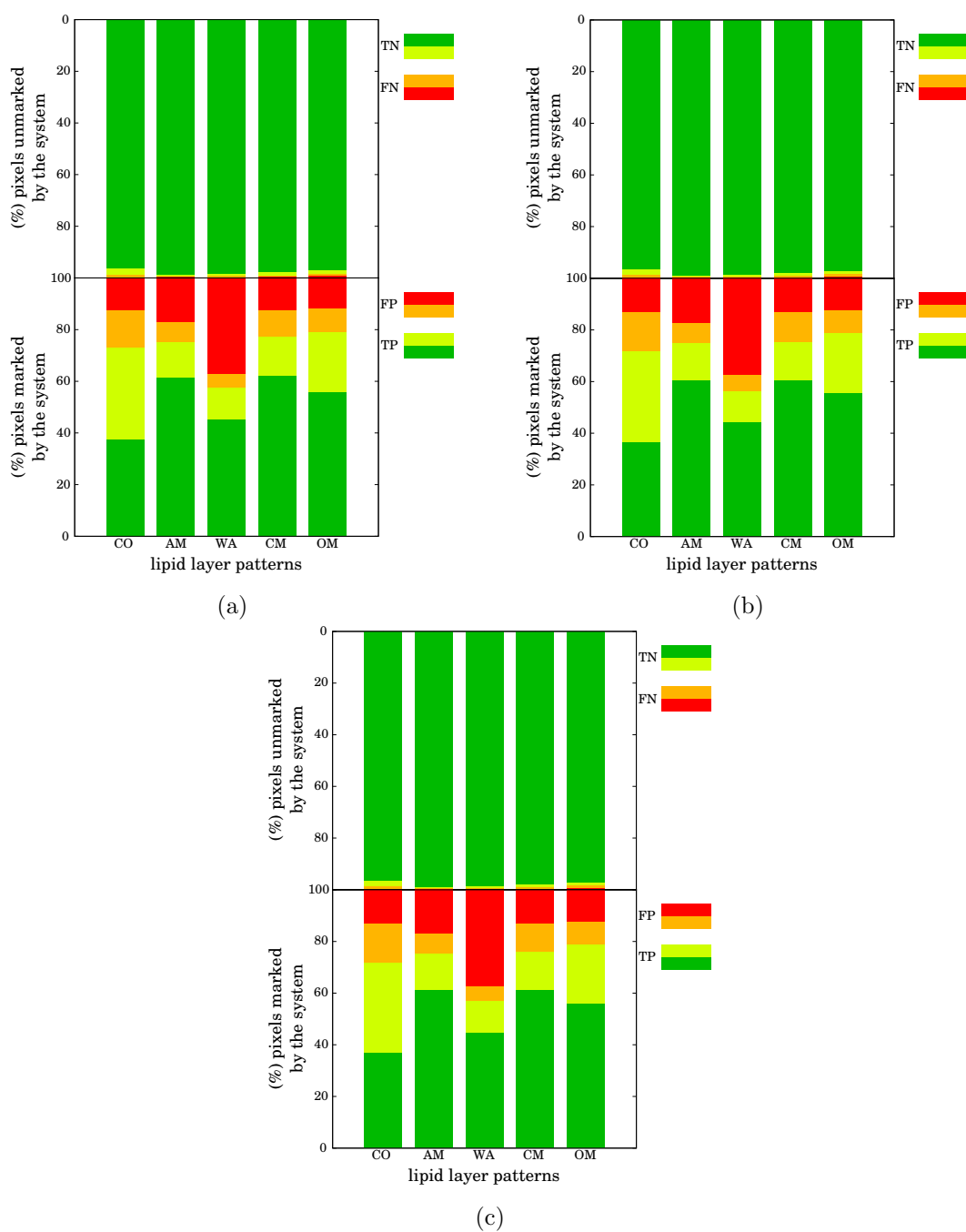


Figure 4.13: Comparison between the system and the three optometrists using the weighted voting system when: (a) $[\omega_1 = 1, \omega_2 = 0, th = 0.9]$, (b) $[\omega_1 = 0, \omega_2 = 1, th = 0.9]$, (c) $[\omega_1 = 1, \omega_2 = 1, th = 0.9]$.

is that, in both terms, the class-membership probabilities of the SVM are taken into account and their values are important enough to be the key of the system.

Table 4.1 presents the performance measures associated to the previous histograms, which confirm the conclusions previously obtained, i.e. there are no significant differences which can be appreciated between these three configurations. Independently of the parameter configuration, the accuracy of the system is over 85% in all the categories, except the wave pattern which seems to be the most difficult one for the system with an accuracy of about 78%. Regarding the sensitivity of the system (TPR), it is quite close to the 100% which means that the system rarely misclassifies those pixels associated to a Guillon category according to 2 or 3 experts. In contrast, the specificity of the system is lower than the sensitivity, which means that the system classifies as Guillon pattern pixels which are not categorized by at least 2 experts. Note that the system is also being penalized by the agreement with only one expert, and so it produces a decrease of both specificity and precision.

Table 4.1: Performance measures using the weighted voting system when: (a) $[\omega_1 = 1, \omega_2 = 0, th = 0.9]$, (b) $[\omega_1 = 0, \omega_2 = 1, th = 0.9]$, (c) $[\omega_1 = 1, \omega_2 = 1, th = 0.9]$.

	Acc	TPR	TNR	Prec		Acc	TPR	TNR	Prec
CO	85.74	97.94	78.49	73.01	CO	85.12	97.97	77.70	71.73
AM	87.27	99.06	80.04	75.24	AM	87.09	99.04	79.82	74.91
WA	78.34	98.49	70.02	57.55	WA	77.78	98.52	69.46	56.40
CM	88.13	98.62	81.36	77.34	CM	87.21	98.70	80.11	75.42
OM	88.76	97.93	82.54	79.21	OM	88.62	97.94	82.33	78.90

(a)

(b)

	Acc	TPR	TNR	Prec
CO	85.23	97.96	77.85	71.97
AM	87.31	99.04	80.11	75.35
WA	78.03	98.52	69.71	56.92
CM	87.60	98.69	80.62	76.21
OM	88.66	97.93	82.39	78.99

(c)

After the analysis focused on weights, the aim now is to explore the impact of the threshold th . Consequently, only one of the above three parameter configurations is considered in combination with two lower thresholds, $th = 0.8$ and $th = 0.7$. In this

way, the regions provided by the system are bigger since the new pixels included have a lower probability of belonging to the corresponding Guillon category. Figure 4.14 depicts the stacked histograms for these two configurations. Some of the new areas detected by the system match with regions marked by the three optometrists, and so the number of false negatives is even lower. However, a lower threshold produces some regions with less reliability and so the number of false positives increases. Regarding the tendency among the different types of patterns, it is exactly the same: the color fringe and the amorphous category are, respectively, the most difficult and the easiest patterns to be categorized by the system. This fact is because the probabilities are provided by the same classifier, and the only difference is the minimum probability used to distinguish the Guillon categories from the background.

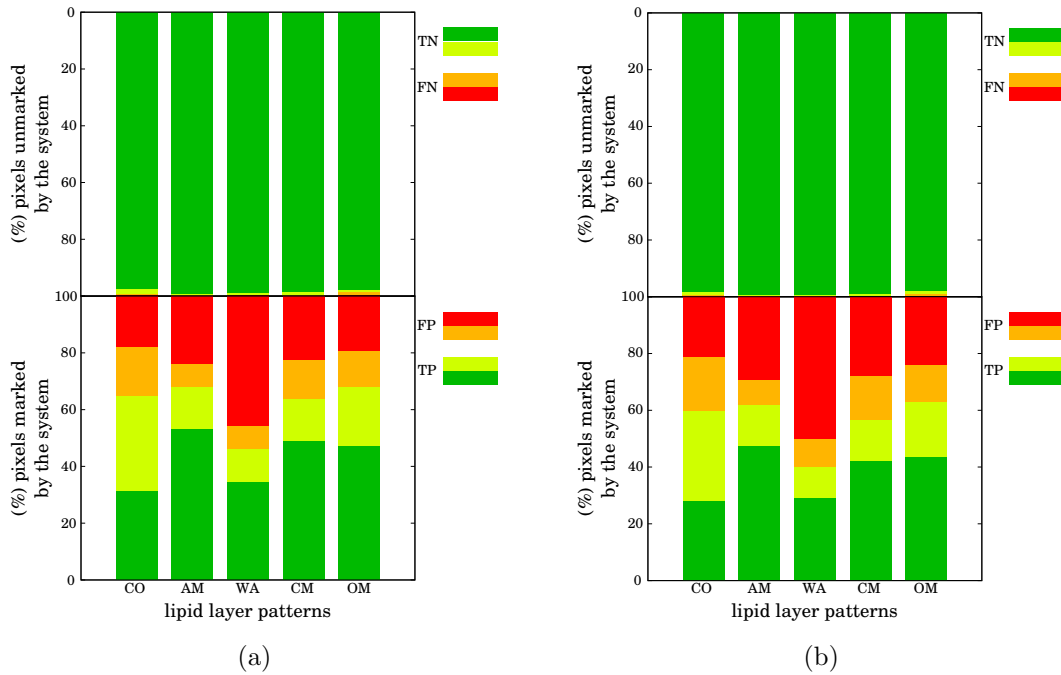


Figure 4.14: Comparison between the system and the three optometrists using the weighted voting system when: (a) $[\omega_1 = 1, \omega_2 = 1, th = 0.8]$, (b) $[\omega_1 = 1, \omega_2 = 1, th = 0.7]$.

Table 4.2 shows the previous results in numerical terms. As expected, the performance measures which take into account the number of false positives have a lower value in this case. The only measure which improves when the threshold is reduced is the specificity due to the decrease in the number of false negatives. Nevertheless, this improvement is slight in comparison with the degradation of the other three performance measures.

Table 4.2: Performance measures using the weighted voting system when: (a) $[\omega_1 = 1, \omega_2 = 1, th = 0.8]$, (b) $[\omega_1 = 1, \omega_2 = 1, th = 0.7]$.

	Acc	TPR	TNR	Prec		Acc	TPR	TNR	Prec
CO	82.00	98.52	73.87	64.97	CO	79.60	98.85	71.24	59.90
AM	83.79	99.31	75.70	68.05	AM	80.86	99.46	72.43	62.06
WA	72.91	98.80	64.97	46.39	WA	69.91	98.94	62.50	40.25
CM	81.66	99.05	73.37	63.92	CM	78.16	99.23	69.72	56.75
OM	83.32	98.12	75.48	67.94	OM	80.98	98.29	72.80	63.05

(a) (b)

Experiment 2: seeded region growing

Using the seeded region growing algorithm, the impact of the β parameter is analyzed. Note that, since square windows are used and they do not have a central pixel, the four central pixels are considered as the window center in both search of seeds and growing steps. The threshold for the search of seeds is $\alpha = 0.9$, which was set empirically since it provides an enough number of seeds. That is, a lower threshold generates a bigger set of seeds but which correspond to the same final regions so they do not imply any improvement, whilst increase the complexity of the procedure. Notice that this threshold is compared with the probability of the classifier, whose maximum value is 1.

Three different representative values where considered for the β parameter: 0.01, 0.05 and 0.1. Figure 4.15 depicts the influence of these three values by means of stacked histograms. Note that the higher the growing threshold, the higher the regions provided by the methodology since the homogeneity criterion of the growing step is less restrictive. That is, the higher the threshold, the higher the number of false positives, and the higher the number of true negatives. The difficulty of this system classifying the Guillon categories is the same than using the weighted voting system. The reason is the classifier and so its outputs are exactly the same, since the difference between the two algorithms lies in the way of using the output probabilities to create the tear film maps. Despite the fact that both methods are based on the same information, the way of using it produces differences in the behavior of the final system. It can be appreciated in the histograms that the size of the green bars are bigger in the seeded region growing algorithm than in the weighted voting system, and so the former performs better and provides more reliable results.

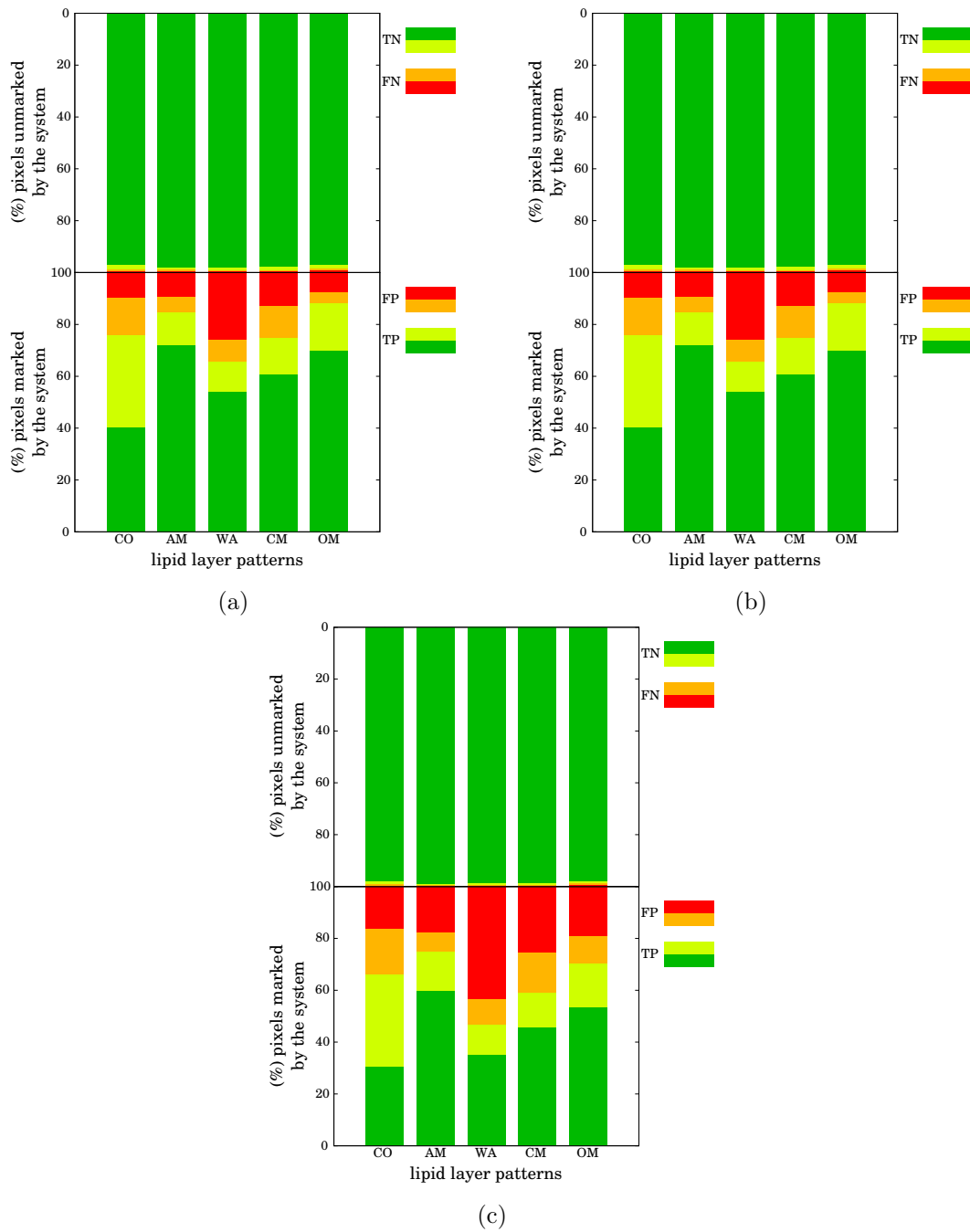


Figure 4.15: Comparison between the system and the three optometrists using the seeded region growing algorithm when: (a) $[\alpha = 0.9, \beta = 0.01]$, (b) $[\alpha = 0.9, \beta = 0.05]$, (c) $[\alpha = 0.9, \beta = 0.1]$.

Table 4.3 shows the performance of the system by means of the four measures considered. As expected, according to the bigger size of the green bars above commented, the accuracy of the system increases for the best configuration of the seeded region growing algorithm, which corresponds to the parameters $[\alpha = 0.9, \beta = 0.01]$. The accuracy of the system is over 80% in all the patterns, and even surpasses the 90% in two of them (amorphous and open meshwork patterns). Not only is the accuracy of the system improved, but also its specificity and precision due to the decrease in the number of false positives. Regarding the sensitivity, its value is slightly lower or higher depending of the Guillon pattern, but with no significant differences. If the analysis is focused on the impact of the β parameter, it can be appreciated that the lower this parameter, the higher the number of false positives. Thus, all the measures which depends on it are also lower, i.e. specificity and precision. A more permissive homogeneity criterion produces bigger regions, in which some pixels are correctly classified. This implies a reduction in the number of false negatives and so a higher sensitivity, although the differences in this case are almost negligible.

Table 4.3: Performance measures using the seeded region growing algorithm when: (a) $[\alpha = 0.9, \beta = 0.01]$, (b) $[\alpha = 0.9, \beta = 0.05]$, (c) $[\alpha = 0.9, \beta = 0.1]$.

	Acc	TPR	TNR	Prec		Acc	TPR	TNR	Prec
CO	87.32	98.18	80.46	76.05	CO	84.23	98.42	76.47	69.58
AM	91.81	98.59	86.68	84.83	AM	89.71	98.88	83.43	80.32
WA	82.44	98.45	74.38	65.91	WA	76.31	98.66	68.03	53.35
CM	86.82	98.57	79.66	74.74	CM	81.56	98.83	73.32	63.88
OM	93.30	98.12	89.35	88.28	OM	85.79	98.05	78.51	73.02

(a)

(b)

	Acc	TPR	TNR	Prec
CO	82.64	98.69	74.55	66.16
AM	87.20	99.02	79.97	75.14
WA	72.98	98.64	65.04	46.60
CM	79.20	98.98	70.80	59.00
OM	88.66	97.93	82.39	78.99

(c)

Experiment 2: processing time

After the quantitative validation focused on the comparison between experts, the aim now is to analyze the processing time in both algorithms. Firstly, the processing time is measured when using the weighted voting system. In this case, the number of windows analyzed over the ROI is always the same since all the windows are processed, pixel by pixel. Secondly, the processing time is measured when using the seeded region growing, which avoids the exhaustive processing of the previous method. On the one hand, not all the windows of the ROI are analyzed in the search of seeds, only those windows separated by at least $\frac{ws}{2}$ are processed, where ws is the window size considered. On the other hand, only the neighbors of those pixels which are part of an existing region are analyzed in the growing process. Furthermore, as the four central pixels of the windows are considered as their centers, instead of doing the growing process pixel to pixel, a total of four pixels are added to a region in each iterative step. In this second case, the impact of the β parameter in the processing time is also analyzed since the higher the growing threshold, the higher the processing time.

Table 4.4 shows a comparative of the times needed to generate a tear film distribution map with both methods. Also, the seeded region growing algorithm is analyzed in depth by quantifying the impact of the β threshold in the processing time. As can be seen, the processing time is decreased by more than half independently of the β value. Furthermore, the time can be reduced more than a 70% in the best case, which corresponds to the minimum β value considered.

Table 4.4: Average time in seconds to create tear film distribution maps.

Algorithm	Time (s)
Weighted voting system	3802.82
$\beta = 0.01$	1091.07
Region growing $\beta = 0.05$	1458.13
$\beta = 0.1$	1620.18

4.3 Conclusions

The previous approach which deals with tear film lipid layer classification provide, as a result, the Guillon category present at the bottom part of the iris according to optometrists' recommendations. However, the patterns defined by Guillon can

appear all around the iris and more than one can be found in a single image. For this reason, a new methodology to create tear film distribution maps was proposed. This methodology uses the previously proposed techniques for color and texture analysis in order to go further in the research. In addition, it includes a soft classification and three different alternatives to create the tear film maps.

The results obtained with this methodology provide the information about the spatial distribution of the different patterns which appear in an input image, and also their location. Regarding the five Guillon categories, the previous approach based on a global classification does not consider images within the amorphous category because it is a very uncommon pattern. The amorphous pattern was considered for the first time in the local approach, with no degradation in performance. Regarding the local analysis, windows of 32×32 pixels were established as the most adequate size for segmentation purposes.

The first attempt to create tear film distribution maps based on the Guillon categories consists of a decision voting system. Also, it tries to take advantage of the global approach and so it uses the unreal background category by means of a hard classification. This preliminary approach produces reliable results in comparison with the annotations done by the optometrists, which demonstrates the feasibility of the problem.

The main disadvantage of the previous approach is the use of the background category, which represents the areas of an image in which no interference pattern has been identified. Since the samples of this unreal category have a high level of variability, they cannot be defined by uniform texture color features and so the accuracy of a classifier may be affected. In order to avoid this problem, a new alternative is proposed based on a weighted voting system. It is focused on two variables: probabilities and distances. This new method takes into account the multiclass probabilities provided by a soft classifier, and a minimum threshold to confirm the identification of the Guillon categories. In this manner, the areas of the image in which there is no pattern are managed in a different way, and the background category is replaced by a minimum threshold. Results obtained with this methodology demonstrate that the tear film maps provided by the developed system are qualitatively similar to the annotations done by three experienced optometrists. Furthermore, the quantitative analysis performed demonstrates that the system produces reliable results with an accuracy over 80% in most cases. Nevertheless, it processes all the windows inside a region of interest and, although the feature extraction time over a single window is almost negligible (under 1 second), analyzing all the windows takes too long (about an hour on average).

So as to reduce the processing time, a last alternative is presented to create tear film distribution maps. The classic seeded region growing algorithm is adapted to the problem at hand, and the class-membership probabilities provided by a soft classifier are used as the homogeneity criterion. This new method is able to generate tear film distribution maps really similar to the regions marked by the optometrists, with a high level of agreement between it and three experienced optometrists. In fact, it noticeably improves previous results in terms of three of the four performance measures considered, with an accuracy over 90% in some cases; with only a slight decrease in the sensitivity. Furthermore, it improves previous approach by a noticeable reduction in the processing time, which decreases over the 70% (from more than 60 minutes to less than 20 minutes).

In clinical terms, the manual process done by optometrists, which consists in localizing each pattern by hand, can be automated with the main benefit of being unaffected by subjective factors. Besides the fact that the system produces unbiased results, it saves time for the experts and provides a detailed distribution of the interference patterns over the input image. In this sense, the experts can have a detailed information of a patient's tear film which means a great help in the diagnosis and treatment of dry eye syndrome.

Chapter 5

Conclusions

Dry eye syndrome is a common clinical condition, whose etiology and management challenge clinicians and researchers alike. It affects a relatively large proportion of the population, and many sufferers will require treatment with a significant potential cost. Monitoring the effect of the different treatments is, therefore, of great importance in ensuring the maximum benefit to each individual.

Its diagnosis is a difficult task due to its multifactorial etiology, and so there exist several clinical tests. One of these tests is the evaluation of the interference patterns of the tear film lipid layer. Guillon designed the Tearscope Plus, an instrument which allows clinicians to rapidly assess the lipid layer thickness, and also defined a grading scale composed of five categories. The classification into these five patterns is a difficult clinical task, especially with thinner lipid layers which lack color and/or morphological features. Therefore, the development of a computer-based analysis is highly desirable, relieving the experts from this tedious task.

Several automated assessments of the tear film lipid layer patterns have been proposed and developed in this PhD thesis. These automated assessments are not intended to override the judgment of an expert in individual cases, but they should prove helpful in the conduct of clinical routine and research.

Initially, a methodology has been presented to assess the tear film lipid layer by automatically classifying images acquired with the Tearscope Plus into the Guillon categories. The process is carried out using texture and color analysis techniques, and machine learning algorithms. The use of color information improves the results compare to grayscale because some lipid layers contain not only morphological features, but also color features. All texture analysis methods perform quite well providing results over the 90% in some cases. In short, the combination of the co-occurrence features analysis and the Lab color space produces the best classification result with maximum accuracy over 96%.

This methodology is able to provide reliable results, but at the expense of a too long processing time and too much memory, since many features have to be computed. This fact makes this methodology unfeasible for practical applications and prevents its clinical use. Consequently, different dimensionality reductions methods are proposed to reduce its computational complexity. This optimization is focused on the improvement of the accuracy and the memory/time requirements. Firstly, the PCA technique has been applied, as a feature extraction method. Its use allows the reduction in memory requirements by transforming the input space and produces no degradation in performance. However, as a transformation is applied, the whole feature vector has to be calculated and so there is no reduction in time. In this manner, feature selection techniques are applied and so, when an input is decided to be unnecessary, the time used in order to calculate it can be saved. Concretely, three of the most popular feature selection filters have been chosen: CFS, consistency-based and INTERACT. They have been tested on the five texture analysis methods considered and the Lab color space. Results obtained with this new step surpass previous results in terms of processing time whilst maintaining accuracy. Additionally, a modification of the ReliefF filter for cost-based feature selection, called mC-ReliefF, has been applied to the problem. The mC-ReliefF allowed to significantly decrease the required time while maintaining the classification performance. Quantitatively, the *ad-hoc* feature selection process based on the CFS filter, which reduces the number of features from 588 to 23 with no degradation in performance, is the one that produces the best balance between accuracy and processing time. It allows the automation of the manual process with maximum accuracy over 97% and processing time under 1 second. Thus, it is completely recommended the use of the proposed methodology for clinical purposes as a supporting tool to diagnose EDE.

Since the heterogeneity of the tear film lipid layer makes its classification into a single category not always possible, tear film maps has been finally presented to illustrate the spatial distribution of the lipid layer patterns. In this manner, more memory and time requirements are needed in exchange for a more detailed information about the localization and size of the patterns over the tear film. Three different approaches has been proposed to tackle the problem: a basic decision voting system, a weighted voting system based on distances and probabilities, and an adapted version of the classic seeded region growing algorithm. The first approach has simply demonstrated the feasibility of the problem, since it provides tear film maps qualitatively similar to the annotations done by the three experienced optometrists. The second alternative is focused on two variables (probabilities and distances), and takes into account the multiclass probabilities provided by a soft classifier. The

quantitative analysis performed demonstrates that the system produces reliable results with an accuracy over 80% in most cases. Nevertheless, it processes all the windows inside a region of interest and, although the feature computing time over a single window is almost negligible (under 1 second), analyzing all the windows takes too long (around an hour on average). Thus, a third and last alternative has been presented. It is based on the classic seeded region growing algorithm and uses the class-membership probabilities provided by a soft classifier as the homogeneity criterion. This method is able to generate tear film distribution maps really similar to the regions marked by the optometrists, with an accuracy over 90% in some cases. In addition, it improves the previous approach by a noticeable reduction in the processing time, which decreases over the 70% (from more than 60 minutes to less than 20 minutes). In summary, tear film distribution maps provide to the experts detailed information of a patient's tear film, which means a great help in the diagnosis and treatment of dry eye syndrome.

5.1 Further research

The proposed methodologies process single images selected by optometrists from a video of the tear film. In this sense, it would be of great interest the investigation of dynamic changes seen in the tear film during the inter-blink time interval. This dynamic analysis could help in identifying subjects with poor tear film stability. In addition, the future lines of research also include the use of alternative algorithms for tear film segmentation. Instead of using the class-membership probabilities associated to texture properties, it could be possible to use directly these properties by means of, for example, edgeless active contours algorithms.

Despite the real-time availability of the system to perform global analysis, there is still large room for improvement on processing time since the local approach takes tens of minutes to provide results. Although the time needed to compute each feature vector is less than 1 second, the great number of features vectors per single image leads us to a huge processing time. As the computing of each feature vector does not depend on calculating any other descriptors, all algorithms proposed for tear film distribution maps can be optimized by means of parallel programming. The idea lies in computing each feature vector, or a set of vectors, in a different processor and finally combining all this information to create the tear film map.

Guillon defined a grading scale composed of the 5 categories previously presented, in such a way that each category represents a range of values of the lipid layer thickness. In order to do this scale more accurate, he also defined some interme-

diate categories. Why not avoiding the specific patterns and proposing a continuous grading scale? In this manner, instead of representing the tear film maps using 5 different colors, which correspond to the 5 categories, the distribution of the lipid layer patterns would be represented by a continuous scale of colors. Thus, each color in the continuous scale would correspond with a specific value of lipid layer thickness, instead of a range of thickness values.

Several devices, based on optical principles, have been designed to assess the lipid layer patterns through the interference phenomena. The Doane interferometer is the instrument employed by the team from the Department of Life Sciences (Glasgow Caledonian University, UK) who have also collaborated in this research. Some experiments were carried out with the interferometry images acquired with this instrument, and promising results were obtained. Thus, the future research also includes the improvement of the methodologies proposed, and as a result a more versatile system would be available for optometrists and practitioners to automatically assess tear film lipid layer patterns using both kind of images.

Appendix A

Experimental results

Some of the experiments presented in both Chapters 2 and 3 are presented here for reasons of legibility. In this sense, the details of these experiments and their results are subsequently described.

A.1 Texture analysis

The target of these experiments is to find which color and texture properties describe better the interference patterns. In this sense, one experiment was carried out per each texture analysis method. Also, an extra experiment was performed with all the possible combinations of texture analysis methods.

Experiment TA1: Butterworth filters for texture analysis

- Texture analysis: Butterworth filters.
- Filters: 9 frequency bands filters.
- Descriptor: 16-bin histograms.
- Color analysis: grayscale, opponent colors and Lab.
- Goal: more appropriate color space and frequency bands.
- Table A.1: from top to bottom, each cell contains the results of grayscale, opponent colors and Lab. The best result per color space appears highlighted.

Experiment TA2: Gabor filters for texture analysis

- Texture analysis: Gabor filters.
- Filters: 16 filters centered at 4 frequencies and 4 orientations.
- Descriptor: 3-bin, 5-bin, 7-bin and 9-bin histograms.
- Color analysis: grayscale, opponent colors and Lab.
- Goal: more appropriate color space and histogram size.
- Table A.2: the best result per color space appears highlighted.

Table A.2: Gabor filters: SVM classification accuracy (%).

	Grayscale	Opponent colors	Lab
3-bin histogram	88.57	86.67	92.38
5-bin histogram	87.62	88.57	94.29
7-bin histogram	86.67	88.57	95.24
9-bin histogram	86.67	88.57	95.24

Experiment TA3: the discrete wavelet transform for texture analysis

- Texture analysis: the discrete wavelet transform.
- Mother wavelets: Haar and Daubechies (Daub4, Daub6, Daub8).
- Number of scales: from 1 to 5.
- Color analysis: grayscale, opponent colors and Lab.
- Goal: more appropriate color space, mother wavelet and number of scales.
- Table A.3: from top to bottom, each cell contains the results of grayscale, opponent colors and Lab. The best result per color space appears highlighted.

Table A.3: The discrete wavelet transform: SVM classification accuracy (%).

	1	2	3	4	5
	79.05	85.71	89.52	87.62	89.52
Haar	84.76	86.67	89.52	90.48	90.48
	83.81	93.33	93.33	92.38	93.33
	79.05	85.71	86.67	87.62	86.67
Daub4	87.62	87.62	89.52	90.48	90.48
	83.81	92.38	91.43	93.33	93.33
	81.91	80.95	85.71	83.81	85.71
Daub6	83.81	86.67	90.48	91.43	91.43
	88.57	90.48	93.33	94.29	94.29
	81.91	85.71	84.76	83.81	84.76
Daub8	84.76	86.67	89.52	89.52	91.43
	85.71	88.57	88.57	91.43	93.33

Experiment TA4: Markov random fields for texture analysis

- Texture analysis: Markov random fields.
- Distances: from 1 to 10.
- Color analysis: grayscale, opponent colors and Lab.
- Goal: more appropriate color space and distance.
- Table A.4: from top to bottom, each cell contains the results of grayscale, opponent colors and Lab. The best result per color space appears highlighted.

Table A.4: Markov random fields: SVM classification accuracy (%); cell ij depicts the results obtained combining the distances ranging from i to j .

	1	2	3	4	5	6	7	8	9	10
	61.91	76.19	81.91	82.86	83.81	82.86	81.91	82.86	85.71	85.71
1	84.76	87.62	88.57	87.62	87.62	86.67	85.71	83.81	84.76	84.76
	66.67	80.95	86.67	87.62	85.71	85.71	84.76	84.76	83.81	84.76

Table A.4: continued from previous page.

	1	2	3	4	5	6	7	8	9	10
		78.10	80.95	81.91	84.76	82.86	83.81	85.71	87.62	85.71
2		80.00	85.71	85.71	87.62	83.81	83.81	82.86	83.81	84.76
		78.10	86.67	84.76	85.71	85.71	84.76	84.76	84.76	84.76
			78.10	81.91	84.76	84.76	84.76	85.71	84.76	84.76
3			84.76	80.95	81.91	82.86	81.91	80.95	80.95	82.86
			83.81	86.67	84.76	81.91	82.86	83.81	84.76	83.81
				83.81	84.76	84.76	81.91	83.81	84.76	82.86
4				82.86	80.00	80.95	81.91	81.91	80.95	81.91
				80.00	81.91	80.95	82.86	82.86	82.86	82.86
					81.91	80.95	79.05	80.95	80.00	80.00
5					80.00	81.91	81.91	82.86	81.91	80.95
					80.95	80.00	80.95	83.81	82.86	82.86
						80.00	77.14	79.05	80.00	80.95
6						84.76	81.91	82.86	81.91	80.00
						80.00	82.86	82.86	82.86	82.86
							75.24	78.10	77.14	77.14
7							77.14	78.10	80.95	78.10
							82.86	81.91	82.86	82.86
								76.19	75.24	78.10
8								80.00	78.10	75.24
								80.00	81.91	78.10
									73.33	77.14
9									76.19	76.19
									79.05	78.10
										77.14
10										73.33
										74.29

Experiment TA5: co-occurrences features for texture analysis

- Texture analysis: co-occurrence features.
- Distances: from 1 to 7.
- Color analysis: grayscale, opponent colors and Lab.
- Goal: more appropriate color space and distance.
- Table A.5: from top to bottom, each cell contains the results of grayscale, opponent colors and Lab. The best result per color space appears highlighted.

Table A.5: Co-occurrence features: SVM classification accuracy (%); cell ij depicts the results obtained combining the distances ranging from i to j .

	1	2	3	4	5	6	7
	80.00	84.76	87.62	90.48	92.38	92.38	92.38
1	85.71	88.57	89.52	89.52	89.52	91.43	91.43
	89.54	91.43	91.43	91.43	94.29	95.24	95.24
		84.76	88.57	89.52	89.52	90.48	91.43
2		89.52	89.52	90.48	89.52	90.48	91.43
		90.48	91.43	92.38	94.29	94.29	94.29
			87.62	90.48	91.43	91.43	92.38
3			90.48	92.38	90.48	90.48	90.48
			94.29	94.29	95.24	94.29	95.24
				89.52	90.48	90.48	92.38
4				89.52	92.38	90.48	89.52
				94.29	93.33	95.24	95.24
					91.43	91.43	90.48
5					90.48	91.43	89.52
					95.24	93.33	95.24
						90.48	91.43
6						90.48	92.38
						96.19	95.24
							92.38
7							91.43
							95.24

Experiment TA6: combination of methods for texture analysis

- Texture analysis: combination of the five methods in all possible ways.
- Parameter configuration: the best result for each pair texture-color.
- Color analysis: grayscale, opponent colors and Lab.
- Goal: increase of the predictive accuracy.
- Tables A.6, A.7, A.8, A.9: the combinations which improve the results of the individual methods appear highlighted.

- Table A.6: combinations of the methods two by two.
- Table A.7: combinations of the methods three by three.
- Table A.8: combinations of the methods four by four.
- Table A.9: combinations of the five methods.

Table A.6: Two method combinations: SVM classification accuracy (%).

	Grayscale		Opponent colours		Lab	
Butterworth filters	83.81	91.43	89.52	91.43	93.33	94.29
Discrete wavelet transform	89.52		91.43		94.29	
Butterworth filters	83.81	91.43	89.52	92.38	93.33	96.19
Co-occurrence features	92.38		92.38		96.19	
Butterworth filters	83.81	90.48	89.52	93.33	93.33	85.71
Markov random fields	83.81		84.76		83.81	
Butterworth filters	83.81	88.57	89.52	87.62	93.33	95.24
Gabor filters	88.57		88.57		95.24	
Discrete wavelet transform	89.52	93.33	91.43	90.48	94.29	98.10
Co-occurrence features	92.38		92.38		96.19	
Discrete wavelet transform	89.52	91.43	91.43	92.38	94.29	94.29
Markov random fields	83.81		84.76		83.81	
Discrete wavelet transform	89.52	93.33	91.43	91.43	94.29	94.29
Gabor filters	88.57		88.57		95.24	
Co-occurrence features	92.38	94.29	92.38	91.43	96.19	96.19
Markov random fields	83.81		84.76		83.81	
Co-occurrence features	92.38	92.38	92.38	92.38	96.19	94.29
Gabor filters	88.57		88.57		95.24	
Markov random fields	83.81	94.29	84.76	91.43	83.81	93.33
Gabor filters	88.57		88.57		95.24	

Table A.7: Three method combinations: SVM classification accuracy (%).

	Grayscale	Opponent colors	Lab
Butterworth filters	83.81	89.52	93.33
Discrete wavelet transform	89.52	94.29	91.43 90.49 94.29
Co-occurrence features	92.38	92.38	96.19
Butterworth filters	83.81	89.52	93.33
Discrete wavelet transform	89.52	93.33	91.43 92.38 94.29 94.29
Markov random fields	83.81	84.76	83.81
Butterworth filters	83.81	89.52	93.33
Discrete wavelet transform	89.52	95.24	91.43 92.38 94.29 94.29
Gabor filters	88.57	88.57	95.24
Butterworth filters	83.81	89.52	93.33
Co-occurrence features	92.38	96.19	92.38 91.43 96.19 96.19
Markov random fields	83.81	84.76	83.81
Butterworth filters	83.81	89.52	93.33
Co-occurrence features	92.38	93.33	92.38 92.38 96.19 94.29
Gabor filters	88.57	88.57	95.24
Butterworth filters	83.81	89.52	93.33
Markov random fields	83.81	95.24	84.76 92.38 83.81 93.33
Gabor filters	88.57	88.57	95.24
Discrete wavelet transform	89.52	91.43	94.29
Co-occurrence features	92.38	95.24	92.38 91.43 96.19 97.14
Markov random fields	83.81	84.76	83.81
Discrete wavelet transform	89.52	91.43	94.29
Co-occurrence features	92.38	94.29	92.38 90.48 96.19 97.14
Gabor filters	88.57	88.57	95.24
Discrete wavelet transform	89.52	91.43	94.29
Markov random fields	83.81	98.10	84.76 93.33 83.81 95.24
Gabor filters	88.57	88.57	95.24
Co-occurrence features	92.38	92.38	96.19
Markov random fields	83.81	96.19	84.76 91.43 83.81 94.29
Gabor filters	88.57	88.57	95.24

Table A.8: Four method combinations: SVM classification accuracy (%).

	Grayscale		Opponent colors		Lab	
Butterworth filters	83.81		89.52		93.33	
Discrete wavelet transform	89.52	95.24	91.43	91.43	94.29	97.14
Co-occurrence features	92.38		92.38		96.19	
Markov random fields	83.81		84.76		83.81	
Butterworth filters	83.81		89.52		93.33	
Discrete wavelet transform	89.52	93.33	91.43	90.48	94.29	97.14
Co-occurrence features	92.38		92.38		96.19	
Gabor filters	88.57		88.57		95.24	
Butterworth filters	83.81		89.52		93.33	
Discrete wavelet transform	89.52	96.19	91.43	94.29	94.29	95.24
Markov random fields	83.81		84.76		83.81	
Gabor filters	88.57		88.57		95.24	
Butterworth filters	83.81		89.52		93.33	
Co-occurrence features	92.38	96.19	92.38	91.43	96.19	94.29
Markov random fields	83.81		84.76		83.81	
Gabor filters	88.57		88.57		95.24	
Discrete wavelet transform	89.52		91.43		94.29	
Co-occurrence features	92.38	95.24	92.38	90.48	96.19	98.10
Markov random fields	83.81		84.76		83.81	
Gabor filters	88.57		88.57		95.24	

Table A.9: Five method combination: SVM classification accuracy (%).

	Grayscale		Opponent colors		Lab	
Butterworth filters	83.81		89.52		93.33	
Discrete wavelet transform	89.52		91.43		94.29	
Co-occurrence features	92.38	95.24	92.38	90.48	96.19	98.10
Markov random fields	83.81		84.76		83.81	
Gabor filters	88.57		88.57		95.24	

A.2 Classification

The target of these experiments is to test the significance of the differences among the predictive accuracies of the five different classifiers. In this manner, one experiment

was carried out per each classifier using the five texture analysis methods and the three color spaces considered.

Experiment C1: Butterworth filters for texture analysis

- Texture analysis: Butterworth filters.
- Filters: 9 frequency bands filters.
- Descriptor: 16-bin histograms.
- Color analysis: grayscale, opponent colors and Lab.
- Goal: more appropriate classifier.
- Table A.10: from top to bottom, each cell contains the results of grayscale, opponent colors and Lab. The best result per color space appears highlighted.
- Table A.11: normality test and p-value.

Table A.10: Butterworth filters: classification accuracy (%).

Classifiers	Frequency bands								
	1	2	3	4	5	6	7	8	9
NB	50.48	59.05	65.71	60.00	59.05	55.24	48.57	46.67	43.81
	59.05	60.95	57.14	57.14	59.05	53.33	50.48	48.57	44.76
	65.71	71.43	79.05	77.14	74.29	70.48	66.67	46.67	44.76
LMT	62.86	53.33	58.10	62.86	64.76	66.67	58.10	54.29	43.81
	58.10	54.29	66.67	76.19	74.29	61.90	64.76	58.10	51.43
	60.95	72.38	77.14	75.24	81.90	73.33	72.38	60.00	57.14
RT	47.62	41.90	54.29	55.24	60.95	65.71	53.33	52.38	32.38
	48.57	53.33	61.90	67.62	53.33	60.00	62.86	58.10	55.24
	48.57	65.71	75.24	75.24	67.62	72.38	67.62	50.48	45.71
RF	42.86	48.57	62.86	60.00	66.67	64.76	60.00	50.48	48.57
	54.29	65.71	68.57	65.71	68.57	69.52	61.90	53.33	57.14
	63.81	76.19	79.05	80.00	75.24	78.10	74.29	61.90	56.19
SVM	61.90	57.14	73.33	72.38	72.38	66.67	68.57	61.90	53.33
	60.00	70.48	82.86	77.14	84.76	74.29	73.33	66.67	61.90
	63.81	80.95	85.71	88.57	89.52	80.00	75.24	64.76	70.48

Table A.11: Butterworth filters: ANOVA results. SS: sum of squared deviations about the mean, df: degrees of freedom, MS: variance.

Grayscale					
Source	SS	df	MS	F	p-value
Between	976.06	4	244.02	3.74	< 0.05
Within	2611.87	40	65.30		
Total	3587.93	44			
Opponent colors					
Source	SS	df	MS	F	p-value
Between	1640.57	4	410.14	8.16	< 0.05
Within	2009.93	40	50.25		
Total	3650.50	44			
Lab					
Source	SS	df	MS	F	p-value
Between	1097.44	4	274.36	2.56	> 0.05
Within	4287.01	40	107.18		
Total	5384.45	44			

Experiment C2: Gabor filters for texture analysis

- Texture analysis: Gabor filters.
- Filters: 16 filters centered at 4 frequencies and 4 orientations.
- Descriptor: 3-bin, 5-bin, 7-bin and 9-bin histograms.
- Color analysis: grayscale, opponent colors and Lab.
- Goal: more appropriate classifier.
- Table A.12: from top to bottom, each cell contains the results of grayscale, opponent colors and Lab. The best result per color space appears highlighted.
- Table A.13: normality test and p-value.

Table A.12: Gabor filters: classification accuracy (%).

Classifiers	Number of bins			
	3	5	7	9
NB	60.00	59.05	58.10	60.00
	62.86	60.00	62.86	64.76
	81.90	82.86	82.86	82.86
LMT	80.95	77.14	74.29	75.24
	70.48	71.43	71.43	79.05
	78.10	81.90	79.05	78.10
RT	67.62	71.43	67.62	68.57
	64.76	65.71	66.67	61.90
	73.33	80.95	68.57	65.71
RF	73.33	66.67	72.38	69.52
	78.10	72.38	70.48	80.00
	78.10	76.19	81.90	75.24
SVM	88.57	87.62	86.67	86.67
	86.67	88.57	88.57	88.57
	92.38	94.29	95.24	95.24

Table A.13: Gabor filters: ANOVA results. SS: sum of squared deviations about the mean, df: degrees of freedom, MS: variance.

Grayscale					
Source	SS	df	MS	F	p-value
Between	1732.62	4	433.15	95.67	< 0.05
Within	67.92	15	4.53		
Total	1800.53	19			
Opponent colors					
Source	SS	df	MS	F	p-value
Between	457.58	3	152.52	13.68	< 0.05
Within	133.81	12	11.15		
Total	591.39	15			

Table A.13: continued from previous page.

	Lab				
Source	SS	df	MS	F	p-value
Between	1071.28	3	357.09	24.58	< 0.05
Within	174.31	12	14.53		
Total	1245.59	15			

Experiment C3: the discrete wavelet transform for texture analysis

- Texture analysis: the discrete wavelet transform.
- Mother wavelet: Daubechies (Daub6).
- Number of scales: from 1 to 5.
- Color analysis: grayscale, opponent colors and Lab.
- Goal: more appropriate classifier.
- Table A.14: from top to bottom, each cell contains the results of grayscale, opponent colors and Lab. The best result per color space appears highlighted.
- Table A.15: normality test and p-value.

Table A.14: The discrete wavelet transform: classification accuracy (%).

Classifiers	Number of scales				
	1	2	3	4	5
NB	64.76	69.52	68.57	69.52	66.67
	61.91	63.81	68.57	70.48	66.67
	68.57	76.19	76.19	73.33	74.29
LMT	62.86	67.62	75.24	71.43	76.19
	70.48	70.48	77.14	75.24	78.10
	79.05	80.00	79.05	80.00	82.86

Table A.14: continued from previous page.

Classifiers	Number of scales				
	1	2	3	4	5
RT	70.48	67.62	69.52	74.29	78.10
	69.52	76.19	66.67	71.43	65.71
	73.33	64.76	77.14	66.67	73.33
RF	74.29	78.10	80.00	82.86	79.05
	80.00	80.96	80.00	76.19	81.91
	82.86	81.91	87.62	88.57	83.81
SVM	81.91	80.95	85.71	83.81	85.71
	83.81	86.67	90.48	91.43	91.43
	88.57	90.48	93.33	94.29	94.29

Table A.15: The discrete wavelet transform: ANOVA results. SS: sum of squared deviations about the mean, df: degrees of freedom, MS: variance.

Grayscale					
Source	SS	df	MS	F	p-value
Between	839.07	4	209.77	15.71	< 0.05
Within	266.98	20	13.35		
Total	1106.05	24			
Opponent colors					
Source	SS	df	MS	F	p-value
Between	1558.44	4	389.61	33.04	< 0.05
Within	235.84	20	11.97		
Total	1794.28	24			
Lab					
Source	SS	df	MS	F	p-value
Between	1460.16	4	365.04	33.72	< 0.05
Within	216.53	20	10.83		
Total	1676.68	24			

Experiment C4: Markov random fields for texture analysis

- Texture analysis: Markov random fields.

- Distances: from 1 to 10.
- Color analysis: grayscale, opponent colors and Lab.
- Goal: more appropriate classifier.
- Table A.16: from top to bottom, each cell contains the results of grayscale, opponent colors and Lab. The best result per color space appears highlighted.
- Table A.17: normality test and p-value.

Table A.16: Markov random fields: classification accuracy (%).

Classifiers	Distances									
	1	2	3	4	5	6	7	8	9	10
NB	38.10	37.14	38.10	36.19	36.19	36.19	35.24	35.24	34.29	33.33
	58.10	38.10	33.33	34.29	37.14	33.33	34.29	33.33	37.14	38.10
	45.71	42.86	39.05	31.43	31.43	30.48	32.38	31.43	30.48	30.48
LMT	51.43	65.71	60.00	55.24	59.05	53.33	60.95	53.33	52.38	64.76
	78.10	66.67	68.57	65.71	67.62	64.76	60.95	60.95	64.76	60.00
	59.05	60.00	67.62	62.86	67.62	60.00	60.00	51.43	49.52	55.24
RT	52.38	55.24	52.38	47.62	54.29	55.24	57.14	52.38	54.29	60.00
	67.62	58.10	57.14	52.38	59.05	52.38	49.52	50.48	50.48	58.10
	42.86	56.19	52.38	57.14	51.43	48.57	52.38	42.86	39.05	57.14
RF	57.14	71.43	63.81	59.05	58.10	60.95	60.00	60.95	60.00	68.57
	80.00	66.67	60.95	59.05	66.67	65.71	61.90	54.29	60.95	54.29
	51.43	59.05	66.67	55.24	60.00	55.24	60.95	61.90	55.24	62.86
SVM	61.90	78.10	78.10	83.81	81.90	80.00	75.24	76.19	73.33	77.14
	84.76	80.00	84.76	82.86	80.00	84.76	77.14	80.00	76.19	73.33
	66.67	78.10	83.81	80.00	80.95	80.00	82.86	80.00	79.05	74.29

Table A.17: Markov random fields: ANOVA results. SS: sum of squared deviations about the mean, df: degrees of freedom, MS: variance.

Grayscale					
Source	SS	df	MS	F	p-value
Between	8574.66	4	2143.67	109.52	< 0.05
Within	880.82	45	19.57		
Total	9455.48	49			

Table A.17: continued from previous page.

Opponent colors					
Source	SS	df	MS	F	p-value
Between	3251.16	3	1083.72	33.36	< 0.05
Within	1169.32	36	32.48		
Total	4420.47	39			
Lab					
Source	SS	df	MS	F	p-value
Between	4352.63	3	1450.88	47.11	< 0.05
Within	1108.76	36	30.8		
Total	5461.39	39			

Experiment C5: co-occurrence features for texture analysis

- Texture analysis: co-occurrence features.
- Distances: from 1 to 7.
- Color analysis: grayscale, opponent colors and Lab.
- Goal: more appropriate classifier.
- Table A.18: from top to bottom, each cell contains the results of grayscale, opponent colors and Lab. The best result per color space appears highlighted.
- Table A.19: normality test and p-value.

Table A.18: Co-occurrence features: classification accuracy (%).

Classifiers	Distances						
	1	2	3	4	5	6	7
NB	68.57	69.52	72.38	75.24	75.24	72.38	70.48
	67.62	73.33	72.38	73.33	74.29	74.29	75.24
	75.24	83.81	81.90	83.81	85.71	86.67	86.67

Table A.18: continued from previous page.

Classifiers	Distances						
	1	2	3	4	5	6	7
LMT	75.24	78.10	76.19	77.14	77.14	80.00	82.86
	70.48	80.95	78.10	81.90	76.19	79.05	80.95
	80.00	82.86	80.00	83.81	86.67	86.67	82.86
RT	74.29	63.81	75.24	74.29	78.10	71.43	71.43
	63.81	65.71	74.29	64.76	71.43	77.14	62.86
	68.57	76.19	76.19	80.95	71.43	71.43	76.19
RF	71.43	74.29	83.81	84.76	82.86	80.00	76.19
	77.14	81.90	73.33	82.86	78.10	82.86	82.86
	83.81	87.62	81.90	88.57	86.67	90.48	83.81
SVM	80.00	84.76	87.62	89.52	91.43	90.48	92.38
	85.71	89.52	90.48	89.52	90.48	90.48	91.43
	89.52	90.48	94.29	94.29	95.24	96.19	95.24

Table A.19: Co-occurrence features: ANOVA results. SS: sum of squared deviations about the mean, df: degrees of freedom, MS: variance.

Grayscale					
Source	SS	df	MS	F	p-value
Between	1165.72	4	291.43	18.29	< 0.05
Within	477.98	30	15.93		
Total	1643.71	34			
Opponent colors					
Source	SS	df	MS	F	p-value
Between	1778.14	4	444.53	31.13	< 0.05
Within	428.42	30	14.28		
Total	2206.56	34			

Table A.19: continued from previous page.

Lab					
Source	SS	df	MS	F	p-value
Between	135.20	4	331.30	29.18	< 0.05
Within	340.63	30	11.35		
Total	1665.83	34			

A.3 Principal component analysis

The target of these experiments is to analyze the impact of using the PCA technique, in terms of the predictive accuracy. In this way, one experiment was carried out per each color space using all the combinations of the texture analysis methods.

Experiment PCA1: grayscale for color analysis

- Texture analysis: combination of the five methods in all possible ways.
- Parameter configuration: the best result for each texture analysis method.
- Color analysis: grayscale.
- Goal: impact of using the PCA technique.
- Table A.20: the column *None* shows the results when PCA was not applied.

Table A.20: PCA using different variances (%) and grayscale images: SVM classification accuracy (%) and number of features.

Texture analysis	Variance (%)						
	None	99	98	97	96	95	90
Discrete wavelet transform	93.33	94.29	94.29	93.33	92.38	92.38	92.38
Co-occurrence features	45	17	13	11	10	9	6
Co-occurrence features	94.29	95.24	94.29	94.29	94.29	94.29	89.52
Markov random fields	44	18	14	11	10	9	6
Discrete wavelet transform	95.24	95.24	94.29	93.33	94.29	94.29	92.38
Co-occurrence features	61	20	15	12	11	9	6
Markov random fields							

Table A.20: continued from previous page.

Texture analysis	Variance (%)						
	None	99	98	97	96	95	90
Co-occurrence features	96.19	95.24	94.29	93.33	94.29	93.33	91.43
Markov random fields	92	23	17	13	11	10	6
Gabor filters							
Discrete wavelet transform							
Co-occurrence features	95.24	97.14	93.33	95.24	95.24	93.33	93.33
Markov random fields	109	24	18	14	12	11	6
Gabor filters							

Experiment PCA1: opponent colors for color analysis

- Texture analysis: combination of the five methods in all possible ways.
- Parameter configuration: the best result for each texture analysis method.
- Color analysis: opponents colors.
- Goal: impact of using the PCA technique.
- Table A.21: the column *None* shows the results when PCA was not applied.

Table A.21: PCA using different variances (%) and opponent colors: SVM categorisation accuracy (%) and number of features

Texture analysis	Variance (%)						
	None	99	98	97	96	95	90
Butterworth filters	87.62	91.43	91.43	91.43	93.33	94.29	92.38
Gabor filters	384	51	35	26	20	16	7
Discrete wavelet transform	92.38	93.33	93.33	92.38	92.38	91.43	90.48
Markov random fields	78	24	19	16	14	12	8
Butterworth filters	92.38	94.29	94.29	93.33	92.38	93.33	93.33
Co-occurrence features	552	58	43	34	28	23	20
Gabor filters							
Butterworth filters	92.38	93.33	94.29	93.33	95.24	94.29	92.38
Markov random fields	396	53	37	28	22	18	8
Gabor filters							

Table A.21: continued from previous page.

Texture analysis	Variance (%)						
	None	99	98	97	96	95	90
Butterworth filters							
Discrete wavelet transform	90.48	94.29	94.29	93.33	93.33	93.33	94.29
Co-occurrence features	618	60	44	35	29	24	12
Gabor filters							

Experiment PCA2: the Lab color space for color analysis

- Texture analysis: combination of the five methods in all possible ways.
- Parameter configuration: the best result for each texture analysis method.
- Color analysis:: Lab.
- Goal: impact of using the PCA technique.
- Table A.22: the column *None* shows the results when PCA was not applied.

Table A.22: PCA using different variances (%) and the Lab color space: SVM categorization accuracy (%) and number of features.

Texture analysis	Variance (%)						
	None	99	98	97	96	95	90
Butterworth filters	94.29	94.29	94.29	94.29	93.33	92.38	93.33
Discrete wavelet transform	210	50	36	27	22	18	9
Discrete wavelet transform	98.10	98.10	99.05	99.05	99.05	98.10	97.14
Co-occurrence features	150	38	29	25	21	19	12
Butterworth filters	94.29	92.38	93.33	93.33	93.33	93.33	93.33
Discrete wavelet transform	246	53	38	29	24	19	10
Markov random fields							
Butterworth filters	96.19	96.19	96.19	96.19	96.19	96.19	96.19
Co-occurrence features	264	57	43	34	28	23	12
Markov random fields							
Butterworth filters							
Discrete wavelet transform	97.14	96.19	96.19	97.14	96.19	96.19	65.24
Co-occurrence features	330	59	44	35	29	24	13
Markov random fields							

Appendix B

Co-occurrence features

Co-occurrence features analysis (Haralick et al., 1973) is a method for texture extraction based on the computation of the conditional joint probabilities of all pairwise combinations of gray levels. The method consists in generating a set of *gray level co-occurrence matrices*, and extracts several statistical measures from their elements. Specifically, a set of 14 statistical measures was proposed in (Haralick et al., 1973).

For reasons of simplicity, the following notation is used:

- N is the number of distinct gray levels in the input image.
- $p(i, j) = P(i, j)/R$ is the (i, j) th entry in a normalized gray level co-occurrence matrix, where R is a normalizing constant.
- $p_x(i) = \sum_{j=1}^N p(i, j)$ is the i th entry in the marginal-probability vector obtained by summing the rows of $p(i, j)$
- $p_y(j) = \sum_{i=1}^N p(i, j)$
- $p_{x+y}(k) = \sum_{i=1}^N \sum_{\substack{j=1 \\ |i+j|=k}}^N p(i, j), k = 2, 3, \dots, 2N$
- $p_{x-y}(k) = \sum_{i=1}^N \sum_{\substack{j=1 \\ |i-j|=k}}^N p(i, j), k = 0, 1, \dots, N - 1$
- $\sum_i = \sum_{i=1}^N$
- $\sum_j = \sum_{j=1}^N$

B.1 Statistical measures

The 14 statistical measures are defined as follows:

- Angular second moment:

$$f_1 = \sum_i \sum_j \{p(i, j)\}^2 \quad (\text{B.1})$$

- Contrast:

$$f_2 = \sum_{n=0}^{N-1} n^2 \left\{ \sum_{\substack{i=1 \\ |i-j|=n}}^N \sum_{j=1}^N p(i, j) \right\} \quad (\text{B.2})$$

- Correlation:

$$f_3 = \frac{\sum_i \sum_j (ij)p(i, j) - \mu_x \mu_y}{\sigma_x \sigma_y} \quad (\text{B.3})$$

where μ_x , μ_y , σ_x and σ_y are the means and standard deviations of p_x and p_y .

- Variance:

$$f_4 = \sum_i \sum_j (i - \mu)^2 p(i, j) \quad (\text{B.4})$$

- Inverse difference moment:

$$f_5 = \sum_i \sum_j \frac{1}{1 + (i - j)^2} p(i, j) \quad (\text{B.5})$$

- Sum average:

$$f_6 = \sum_{i=2}^{2N} i p_{x+y}(i) \quad (\text{B.6})$$

- Sum variance:

$$f_7 = \sum_{i=2}^{2N} (i - f_6)^2 p_{x+y}(i) \quad (\text{B.7})$$

- Sum entropy:

$$f_8 = - \sum_{i=2}^{2N} p_{x+y}(i) \log\{p_{x+y}(i)\} \quad (\text{B.8})$$

- Entropy:

$$f_9 = - \sum_i \sum_j p(i, j) \log(p(i, j)) \quad (\text{B.9})$$

- Difference variance:

$$f_{10} = \text{variance of } p_{x-y} \quad (\text{B.10})$$

- Difference entropy:

$$f_{11} = - \sum_{i=0}^{N-1} p_{x-y}(i) \log\{p_{x-y}(i)\} \quad (\text{B.11})$$

- Information measures of correlation:

$$f_{12} = \frac{HXY - HXY1}{\max\{HX, HY\}} \quad (\text{B.12})$$

$$f_{13} = (1 - \exp[-2(HXY2 - HXY)])^{1/2} \quad (\text{B.13})$$

where HX and HY are the entropies of p_x and p_y , respectively, and:

$$HXY = - \sum_i \sum_j p(i, j) \log(p(i, j))$$

$$HXY1 = - \sum_i \sum_j p(i, j) \log\{p_x(i)p_y(j)\}$$

$$HXY2 = - \sum_i \sum_j p_x(i)p_y(j) \log\{p_x(i)p_y(j)\}$$

- Maximal correlation coefficient:

$$f_{14} = (\text{Second largest eigenvalue of } Q)^{1/2} \quad (\text{B.14})$$

where:

$$Q(i, j) = \sum_k \frac{p(i, k)p(j, k)}{p_x(i)p_y(j)}$$

Appendix C

Estimating the accuracy of classifiers

This appendix is concerned with estimating the performance of any machine learning algorithm. The most obvious criterion to estimate the performance of a classifier is *accuracy* (Bramer, 2007), which represents the proportion of a set of unseen samples that it properly classifies.

In many cases, the number of possible unseen samples is potentially very large and so it is not possible to establish the accuracy beyond dispute. Instead, it is very common to *estimate* the accuracy of a classifier by measuring its accuracy for a sample of data not used during the training process. There are three popular strategies used for this issue (Bramer, 2007): dividing the data into a training set and a test set, k -fold cross-validation, and N -fold (or leave-one-out) cross-validation. Due to the size of the datasets used in this thesis, only the k -fold and the leave-one-out cross-validations were used, and so they are subsequently described.

C.1 k -fold cross-validation

The k -fold cross-validation (Rodriguez, Perez, & Lozano, 2010) is often used when the number of instances is small, although many researchers make use of it despite size. The process consist in dividing the dataset composed of N samples into k equal parts, where k is typically a small number (5 or 10). Then, an iterative process is carried out k times. At each iteration, one of the k parts is used as a test set, and the remaining $k - 1$ parts are used as a training set.

Finally, the total number of samples correctly classified, in all k iterations, is divided by the total number of samples N to obtain an overall level of accuracy p .

Note that the standard error is $\sqrt{p(1-p)/N}$.

C.2 Leave-one-out cross-validation

The N -fold cross-validation, often known as “leave-one-out” cross-validation, is a particular case of the k -fold cross-validation where the dataset is divided into as many parts as instances (Bramer, 2007). In this manner, N classifiers are generated by training $N - 1$ samples, and each of them is used to classify a single test instance. The predictive accuracy p is the total number of correctly classified instances divided by the total number of them, and so the standard error is $\sqrt{p(1-p)/N}$.

The large amount of computation involved makes this method unsuitable for large datasets. In fact, it is appropriate to be used with very small datasets where as much data as possible needs to be used in order to train the classifier.

Appendix D

Comparing classifiers: statistical analysis

There is no infallible way of finding the best machine learning for a particular problem. One of the possible manners to deal with this issue is to compare the performance of a set of machine learning algorithms, applied over a range of datasets, by performing a statistical analysis in order to find significant differences.

If there are only two classifiers to compare, the mean error/accuracy can be compared by means of the *paired t-test* (Goulden, 1956) or the *Wilcoxon test* (Wilcoxon, 1945). Nonetheless, if the number of algorithms is three or more, it is not appropriate to compare each pair of models using these tests. The reason is that the likelihood of incorrectly detecting a significant difference increases with the number of comparisons. In this case, the proposed methodology to compare a set of classifiers is defined according to Figure D.1. As the ANOVA test can only be applied if the data are normally distributed, the Lilliefors test is firstly applied and its null hypothesis is checked. Thus, if it rejects the null hypothesis, that the data are from a normal distribution, then the comparison of classifiers cannot be performed. Otherwise, the ANOVA test is applied in order to identify if there is a significant difference between all the means. If it accepts the null hypothesis, that all population means are equal, then the simplest classifier is selected. Otherwise, the Tukey's method is applied, a multiple comparison procedure that tests all means pairwise to determine which ones are significantly different.

The statistical methods above mentioned are subsequently explained in depth.

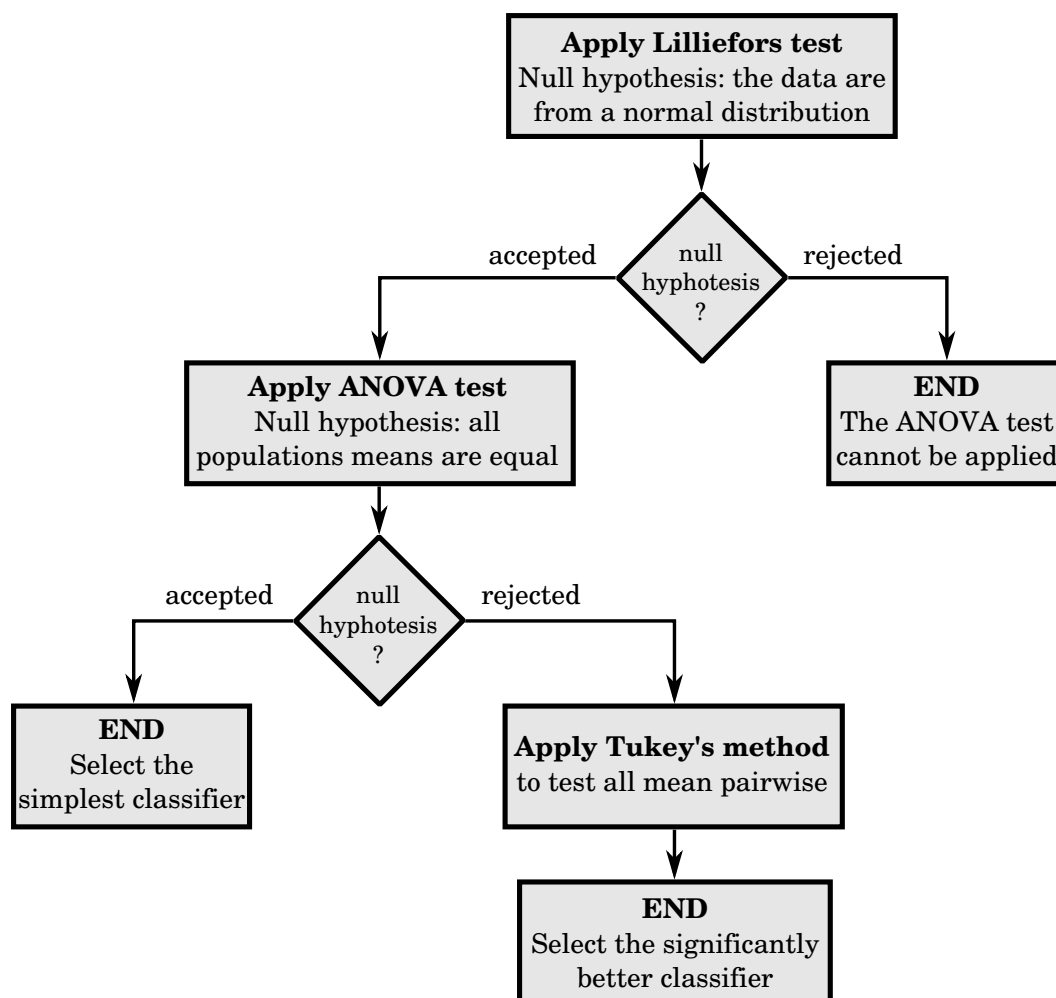


Figure D.1: Steps of the methodology to statistically compare a set of classifiers.

D.1 The Lilliefors test for normality

The normality assumption is at the core of a majority of standard statistical procedures. Among the many procedures used to test this assumption, two of them should be highlighted: the Kolmogorov-Smirnov test (Massey, 1951) and the chi-square test (Moore, 1976). The former one has two main advantages compared to the second one according to (Massey, 1951): it can be used with small sample size, and it is more powerful for any sample size. Regrettably, the Kolmogorov-Smirnov test no longer applies when certain parameters of the distribution must be estimated from the sample.

In this sense, the Lilliefors test (Lilliefors, 1967) was developed based on the Kolmogorov-Smirnov test. It aims at testing if a set of observations come from a normal distribution or not, when the mean and the variance of the distribution are not specified. The procedure is defined as follows:

1. Estimate the mean and the variance of the distribution based on the data.
2. Find the maximum discrepancy between the empirical distribution function, and the cumulative distribution function of the normal distribution with the mean and the variance previously estimated.
3. If the maximum discrepancy exceeds the critical value, then the null hypothesis that the observations are from a normal distribution is rejected. Otherwise, the null hypothesis is accepted.

D.2 The ANOVA test

The ANOVA test (Hogg & Ledolter, 1987) is a statistical test for heterogeneity of means by analysis of group variances. That is to say, it aims at finding out if there are any significant differences among three or more population means. The data must be normally distributed or nearly to apply this test, and so a normality test has to be previous applied. The procedure is defined as follows:

1. Calculate the ANOVA table by comparing the means of several distributions, and estimating the variances among distributions and within a distribution.
2. Compute the p-value from the ANOVA table.
3. If the p-value exceeds the critical value, the null hypothesis that all population means are equal (from the same population or from different populations but with the same mean) is accepted. Otherwise, the null hypothesis is rejected.

Notice that if the p-value does not exceed the critical value, it does not imply that every mean differs from every other mean. It only implies that at least one mean differs from the rest of them.

D.3 The Tukey's method for multiple comparison

There are different methods for multiple comparisons, and most of them are for pairwise of group means. Their target is to determine which group of means are

significantly different from which others. The Tukey's method (Hsu, 1996) is one of the most popular techniques to perform multiple comparisons, and it is used when less conservative test is desirable, i.e., more powerful.

Two main assumptions have to be verified before applying the Tukey's method: (1) the populations are normally distributed, which is tested here with the Lilliefors test; and (2) a decision to reject the null hypothesis that all the means are equal, which is made during the ANOVA test. The procedure is defined as follows:

1. Calculate the test statistic:

$$q = \frac{\bar{x}_j - \bar{x}_i}{\sqrt{\frac{s^2}{2} \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}} \quad (\text{D.1})$$

where $\bar{x}_j > \bar{x}_i$, s^2 is the mean square error estimate of σ^2 from the ANOVA test, and n_i and n_j are the sample sizes from population i and j , respectively.

2. If the q-value exceeds the critical value, then the null hypothesis that the means of populations i and j are equal is rejected. Otherwise, the null hypothesis is accepted and so there is not sufficient evidence to conclude that the means of populations i and j are significantly different.
3. If the null hypothesis is rejected, then state the conclusion of the test based on the decision made and with respect to the pairwise claim.

Appendix E

Evaluation of tear film lipid layer classification

The diagnosis of the dry eye syndrome is complicated since it has no single characteristic sign or symptom, and no single diagnostic measure. There are several clinical tests which can be performed as part of a routine eye care examination. One of them is called *lipid layer pattern assessment*, whereby tear film quality and lipid layer thickness can be assessed by non-invasively imaging the superficial lipid layer by interferometry. This test is based on a standard classification defined by Guillon (Guillon, 1998), who specified various types of lipid layer patterns.

First attempts to automatize tear film lipid layer classification can be found in (Ramos et al., 2011; García-Resúa et al., 2013), where it was demonstrated how the interference phenomena can be characterized as a color texture pattern. These results were later improved in (Remeseiro et al., 2011) by using a set of texture analysis techniques and color spaces, and extended in (Remeseiro et al., 2012) to five different machine learning algorithms. The problem with these approaches, which prevented their clinical use, is that they required a large amount of time for the computation of the features. In (Bolon-Canedo et al., 2012; Remeseiro, Bolon-Canedo, et al., 2014), several feature selection filters were successfully used in order to reduce the number of features for classification and so the time needed for processing. More particularly, a method based on CFS was proposed so that the time was reduced to just under 1 second. Finally, a more systematic procedure for automatic tear film lipid layer classification was proposed in (Méndez, Remeseiro, Peteiro-Barral, & Penedo, 2013). In that research, class-binarization techniques, feature selection methods, and artificial neural networks were used to improve classification performance. Furthermore, for the first time, several performance measures were introduced for tear film classifi-

cation, which were evaluated using TOPSIS as a multiple-criteria decision-making (MCDM) method.

This last work was extended to a more generalizable methodology in (Peteiro-Barral, Remeseiro, Penedo, & Méndez, n.d.). This appendix presents the whole methodology, and is organized as follows. Section E.1 describes the proposed methodology as a pipeline of processes for optimizing and evaluating different solutions to a problem. Section E.2 presents a case of study of the methodology previously proposed to the particular case of automatic tear film lipid layer classification. Section E.3 shows the results of applying the methodology to tear film classification. Finally, Section E.4 shows the conclusions.

E.1 Methodology

This methodology can be used as a baseline in any classification problem to provide several solutions and evaluate their performance (see Figure E.1). The first step entails the data acquisition of the particular problem. Next, the obtained dataset is converted into new datasets through an optimization process that includes class binarization techniques, which may improve performance of the classifiers, and feature selection methods, which may reduce the complexity of the problem. Then, the classification step is performed by means of machine learning algorithms. Finally, all the solutions are evaluated based on their performance measures. For this task, decision-making methods are used to obtain ranking lists of alternatives. Since there can be disagreements between these methods, the conflict handling step provides a solution to obtain a single ranking.

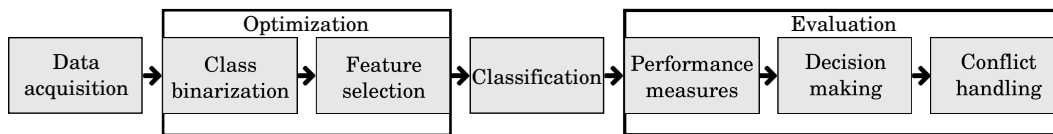


Figure E.1: Steps of the research methodology.

1. *Data acquisition.* It is the sampling of the real world to generate data that can be manipulated by a computer such as temperature, pressure, flow, humidity or other measures. Thus, the result of this stage is a dataset composed of samples which belong to different classes and are represented by features.
2. *Class binarization.* Several machine learning algorithms are inherently designed for binary classification. A class binarization is a mapping of a multi-class learning problem to several two-class learning problems in a way that

allows a sensible decoding of the prediction (Furnkranz, 2003). Moreover, there also exists evidence that even the “single machine” approaches, which construct a multi-class classifier by solving a single optimization problem, may improve performance via class binarization (Dietterich & Bakiri, 1995a). In this way, the dataset obtained in the previous step is transformed to several datasets whose samples belong to only two classes.

3. *Feature selection.* It is the process of selecting a subset of features occurring in the data and using only this subset as features in training and classification, so that the feature space is optimally reduced according to a certain evaluation criterion. Feature selection serves two purposes (Manning, Raghavan, & Schütze, 2008): it makes training more efficient by decreasing the size of the data, and it increases classification accuracy by eliminating noise features and avoiding overfitting (Loughrey & Cunningham, 2005). As a result of this step, samples are represented by a subset of the original features.
4. *Classification.* Supervised learning entails learning a mapping between a set of input features and output labels, and applying this mapping to predict the outputs for new data (Mitchell, 1997). The resulting classifier is then used to assign class labels to the new instances whose values of the features are known, but the value of the class label is unknown (Kotsiantis, 2007). This stage results in a set of classifiers trained with the previously obtained datasets.
5. *Performance measures.* They quantify the behavior of the classification process. The performance of machine learning algorithms for classification is typically evaluated by several measures obtained from a confusion matrix. Multi-class performance is usually evaluated by averaging the individual per-class performance measures. However, this method may be problematic in cases where substantial differences exist across classes because averaging hides details. For this reason, the worst of the individual per-class performance measures as a lower bound estimation procedure (Fernandez-Caballero, Martínez, Hervás, & Gutiérrez, 2010) is considered. In addition to these measures, other works take into account time, model complexity, etc. This step calculates a set of performance values from each trained classifier.
6. *Decision-making.* Real world problems usually consider several performance measures. A multi-criteria problem is formulated using a set of alternatives and criteria, and a decision matrix where x_{ij} is the performance measure of the i -th alternative in the j -th criterion. In the decision-making process criteria are identified, weights are given to each criterion to reflect its relative

importance, and weighted preference scores are derived based on the criteria weights and criteria score. The ultimate result of this process is a ranking list of alternatives.

7. *Conflict handling.* If several decision-making processes are used, then they can offer conflicting rankings of alternatives. Conflict handling is the process by which conflicting rankings are merged into a single ranking. Different methods can be devised to choose an approximate solution to these conflicts from the straightforward average of the rankings to more sophisticated methods (Peng, Kou, Wang, & Shi, 2011). Thus, if there are disagreements between rankings, this step provides a single ranking list of all the alternatives.

E.2 A case of study

The proposed methodology is applied to evaluate tear film lipid layer classification. For this reason, every step of the general methodology is adapted to the problem at hand (see Figure E.2).

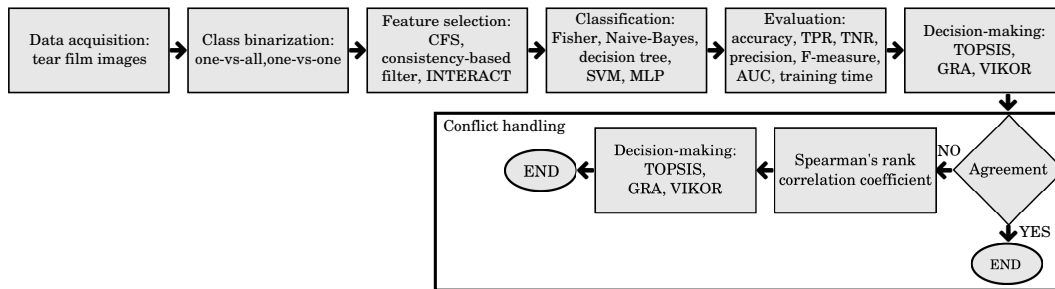


Figure E.2: Steps of the research methodology applied to tear film lipid layer classification.

E.2.1 Data acquisition: tear film images

The steps of data acquisition in the problem of tear film lipid layer classification are (Remeseiro et al., 2011):

1. *Image acquisition.* Input images were captured using the Tearscope plus (Tearscope Plus, 1997), and they were stored at a spatial resolution of 1024×768 pixels in RGB.
2. *Extraction of the region of interest.* Experts that analyze interference images focus their attention on the bottom part of the iris, in which the tear can be perceived with higher contrast. Thus, tear film classification takes place in

this area called the *region of interest* (ROI) and selected according to (Calvo et al., 2010).

3. *Color analysis.* Color information is extracted from the ROIs by using the Lab color space (McLaren, 1976). In order to analyze the texture using Lab, the texture of each component is analyzed individually and three descriptors per image are created. The final descriptor is the concatenation of them.
4. *Texture analysis.* Texture information is extracted from the ROIs by applying the co-occurrence features technique (Haralick et al., 1973), which describes textures as statistical measures. A set of 28 features composes the texture descriptor for a particular distance.

Distances from 1 to 7 in the co-occurrence features method and 3 components of color are considered, so the size of the final descriptor obtained from an input image is: $28 \text{ features} \times 7 \text{ distances} \times 3 \text{ components} = 588 \text{ features}$.

E.2.2 Class binarization techniques

The most common strategies for class binarization are the *one-vs-all* and the *one-vs-one* decompositions, described as follows:

- The *one-vs-all* technique divides a c -class problem into c binary problems. Each problem is solved by a binary classifier which has to distinguish one of the classes from all other classes.
- The *one-vs-one* technique divides a c -class problem into $\frac{c(c-1)}{2}$ binary problems. Each problem is solved by a binary classifier which has to distinguish between a pair of classes.

Once the classifiers are trained, there is the need of decoding methods in order to obtain their outputs. If the algorithms are soft, they compute the “likelihood” of classes for a given input. That is, they obtain a confidence p for the *positive* class and a confidence $1 - p$ for the *negative* class. In the *one-vs-all* technique, if we assume the *one*-part as the positive class and the *all*-part as the negative class, the decoding method is done according to the maximum probability p among classes. However, this method is not valid for *one-vs-one* techniques. Consequently, three different decoding methods for *one-vs-one* binarization techniques are considered:

- *Hamming decoding* (Dietterich & Bakiri, 1995b). This method uses a matrix $M \in \{-1, 1\}^{N \times F}$, where N is the number of classes and F is the number of

binary classifiers. It induces a partition of the classes into two “metaclasses”, where a sample is placed in the positive metaclass for the j -th classifier if and only if $M_{y_i,j} = 1$, where y_i stands for the desired class of the sample.

- *Loss-based decoding* (Allwein, Schapire, & Singer, 2001). The use of the loss function L instead of the Hamming distance is suggested in order to take into account the significance of the predictions, which can be interpreted as a measure of confidence. In this research, the most appropriate loss function is the logistic regression $L(z) = \log(1 + e^{-2z})$.
- *Accumulative probability with threshold* (Allwein et al., 2001). It extends the Hamming matrix to $M \in \{-1, 0, 1\}^{N \times F}$. It ignores binary classifiers if the difference between the confidence for the positive and negative classes is under a threshold.

E.2.3 Feature selection: filters

Feature selection techniques can be divided into three groups: filters, wrappers and embedded methods (Guyon et al., 2006). Both wrappers and embedded methods have the risk of overfitting when having more features than samples (Loughrey & Cunningham, 2005), as in this research. Consequently, filters were chosen since they allow for reducing the dimensionality of the data without compromising the time and memory requirements of machine learning algorithms.

The following three filters were chosen based on previous researches (Bolon-Canedo, Sánchez-Marroño, & Alonso-Betanzos, 2011):

- *Correlation-based feature selection* (CFS) (M. A. Hall, 1999). It is a multivariate filter that ranks feature subsets according to a correlation based heuristic evaluation function. The bias of this function is toward subsets that contain features which are highly correlated with the class and uncorrelated with each other.
- *Consistency-based filter* (Dash & Liu, 2003). This algorithm evaluates the worth of a subset of features by the level of consistency in the class values when the samples are projected onto the subset of attributes.
- *INTERACT* (Zhao & Liu, 2007). It is a subset filter based on symmetrical uncertainty, which is defined as the ratio between the information gain and the entropy of two features. It also includes the consistency contribution of a feature, which is an indicator about how the elimination of that feature will affect consistency.

E.2.4 Classification: machine learning algorithms

Five popular machine learning algorithms were selected aiming to provide different approaches of the learning process:

- *Fisher's linear discriminant* (J. H. Friedman, 1989). It is a simple method used to find the linear combination of features which best separate two or more classes.
- *Naive-Bayes* (Jensen, 1996). It is an statistical learning algorithm based on the Bayesian theorem which can predict class membership probabilities.
- *Decision tree* (Murthy, 1998). It is a logic-based algorithm which classifies samples by sorting them based on feature values.
- *Support vector machine* (SVM) (Burges, 1998). It is based on the statistical learning theory and revolves around a hyperplane that separates two classes.
- *Multilayer perceptron* (MLP) (Rosenblatt, 1958). It is a feedforward artificial neural network which consists of a set of units, joined together in a pattern of connections.

E.2.5 Performance measures

The quality of the results provided by the classifiers are evaluated in terms of the following performance measures:

- *Accuracy*. The percentage of correctly classified instances.
- *True positive rate* (TPR). The proportion of positives which are correctly classified, also called sensitivity or recall.
- *True negative rate* (TNR). The proportion of negatives which are correctly classified, also called specificity.
- *Precision*. The proportion of the true positives against all the positive results.
- *F-measure*. The harmonic mean of precision and recall.
- *Area under the curve* (AUC). The area under the receiver operating characteristic (ROC) curve.

The training time of the algorithms is also considered. The training step is executed off-line so its value is not as relevant as the other measures, but it may be helpful to select the best classifier when other measures are quite similar.

- The *training time* comprises the time elapsed for training a learning model. Notice that this comprises training a set of classifiers when class binarization techniques are used.

Note also that the testing time, that is the time elapsed for outputting a new classification, is negligible thus it will not be considered as a selection criterion.

E.2.6 Decision-making: multiple-criteria decision-making methods

This section gives an overview of the three multiple-criteria decision-making (MCDM) methods that will be used to analyze all the performance measures.

TOPSIS

Technique for order preference by similarity to ideal solution (TOPSIS) (Hwang & Yoon, 1981) is based on the idea of finding the best alternatives by minimizing the distance to the ideal solution whilst maximizing the distance to the negative-ideal solution. The extension proposed in (Opricovic & Tzeng, 2004) is adopted in this research, and involves next steps:

1. Compute the normalized decision matrix.
2. Determine the weights and compute the weighted normalized decision matrix.
3. Determine the best ideal and the worst negative-ideal solutions of all criteria.
4. Compute the coefficient R which measures the relative distance to the ideal and negative-ideal solutions.
5. Rank the alternatives by maximizing the coefficient R .

GRA

Gray relational analysis (GRA) (Kuo, Yang, & Huang, 2008) is based on the degree of similarity or difference of development trends between an alternative and the ideal alternative. The steps involved in GRA are:

1. Calculate the gray relation values.
2. Calculate the ideal solution.
3. Compute the gray relational coefficient between the ideal values and the gray relation values.

4. Compute the gray relational grade Γ , which indicates the closeness between the ideal solution and the alternatives.
5. Rank the alternatives by maximizing the coefficient Γ .

VIKOR

VIKOR (Opricovic, 1998) is a method which provides maximum group utility for the majority and minimum individual regret for the opponent. The procedure used is as follows:

1. Determine the best and the worst values of all criteria.
2. Compute the distance of the alternatives to the ideal and the negative-ideal solutions.
3. Compute the VIKOR coefficient Q of the alternatives.
4. Rank the alternatives by maximizing the coefficient Q .

E.2.7 Conflict handling: Spearman's rank correlation coefficient

Since several MCDM methods are used, differences among the rankings may appear. Thus, the Spearman's rank correlation coefficient is used to handle conflicting MCDM rankings. A weight is assigned to each MCDM method according to the similarities between every pair of rankings generated by the MCDM methods.

The Spearman's rank correlation is a nonparametric technique for evaluating the degree of linear association or correlation between two independent variables (Gauthier, 2001). It is calculated according to the following equation:

$$\rho = 1 - \frac{6 \sum_{i=1}^m d_i^2}{m(m^2 - 1)} \quad (\text{E.1})$$

where d_i is the difference between ranks for each (x_i, y_i) data pair, and m is the number of data pairs.

The procedure adopted for conflict handling is as follows:

1. Compute the average similarities between the k -th method and the other MCDM methods as:

$$\rho_k = \frac{1}{q-1} \sum_{i=1, i \neq k}^q \rho_{ki}, k = 1, 2, \dots, q \quad (\text{E.2})$$

where q is the number of MCDM methods, and ρ_{ki} the Spearman’s rank correlation coefficient between the k -th and i -th MCDM methods.

2. Normalize the ρ_k values $\sum_{k=1}^q \rho_k = 1$ to compute secondary rankings of classifiers, then they can be used as weights for the MCDM methods.
3. Apply the MCDM methods to re-rank all the alternatives using ranking scores produced previously by MCDM methods and the weights obtained by normalizing the Spearman’s rank correlation coefficients.

E.3 Experimental results

This section presents the evaluation of tear film lipid layer classification using the proposed methodology and the specific techniques presented in previous section. Experimentation was performed on Matlab in an Intel[®] Core[™] i5-650 CPU @ 4M Cache, 3.20 GHz with RAM 6GB DDR3. The dataset of images used for validation are those in the VOPTICAL_I1 dataset (VOPTICAL_I1, n.d.).

Regarding the parameters of the classifiers, a SVM with radial basis kernel and automatic parameter estimation, and a MLP with a single hidden layer and a number of hidden units selected according to (Méndez et al., 2013) were considered. Moreover, a leave-one-out cross-validation (see Appendix C) was used to analyze the generalization of the results to larger datasets. Finally, the weights of the measures in the MCDM methods are assigned equally, except for the training time that is reduced to 0.01. The training time is a cost criteria while the other measures are benefit criteria.

E.3.1 Results

Table E.1 shows the number of features selected for every method and feature selection filter used in this research. Also, the percentage of features selected in the total of 588 features is also shown. Notice that the feature selection step is done for every classifier since binarization methods change the output search space. Consistency-based filter selects the smallest subset of features. In average, it retains the 0.54% of the features. Conversely, CFS selects eight times more features (4.47%) than the former. Halfway, INTERACT selects the 3.23% of the features. In terms of class binarization, feature selection retains 3.06% of the features in the single approach, 2.30% in the one-versus-all, and 3.00% in the one-versus-one.

As expected, the percentage of features selected in the single machine, multi-class, method is larger than the percentage of features in class binarization because

this reduces the complexity of the problem. However, the number of features in one-versus-one is larger than in one-versus-all although the former is, a priori, an easier task than the latter. Note that in one-versus-one the number of features selected might increase because of the lack of relevant knowledge in a smaller data set corresponding only with the two classes involved. A smaller number of samples in a 105-sample data set worsens this issue.

Table E.1: Number and percentage of features selected for every method and feature selection filter in the total of 588 features.

Method		CFS	Consistency	INTERACT
Single	—	27 (4.59%)	6 (1.02%)	21 (3.57%)
One- <i>vs</i> -all	1- <i>vs</i> -all	17 (2.89%)	2 (0.34%)	14 (2.38%)
	2- <i>vs</i> -all	27 (4.59%)	6 (1.02%)	17 (2.89%)
	3- <i>vs</i> -all	11 (1.87%)	3 (0.51%)	14 (2.38%)
	4- <i>vs</i> -all	33 (5.61%)	4 (0.68%)	14 (2.38%)
One- <i>vs</i> -one	1- <i>vs</i> -2	20 (3.40%)	2 (0.34%)	12 (2.04%)
	1- <i>vs</i> -3	53 (9.01%)	1 (0.17%)	53 (9.01%)
	1- <i>vs</i> -4	23 (3.91%)	1 (0.17%)	23 (3.91%)
	2- <i>vs</i> -3	27 (4.59%)	3 (0.51%)	14 (2.38%)
	2- <i>vs</i> -4	24 (4.08%)	3 (0.51%)	14 (2.38%)
	3- <i>vs</i> -4	27 (4.59%)	4 (0.68%)	13 (2.21%)

Tables E.2, E.3 and E.4 show the results for every decision-making method. For purposes of simplicity and clarity, only the top 10 results will be shown. Note that the wide set of techniques used in this research define a 96-alternative configurations in total. This table represents the rank and value determined by the corresponding decision-making method; the class binarization method, classifier and feature selection method used; and the performance measures utilized in this research.

As can be seen, the class binarization method one-versus-one, in its different configurations, populates 60% of the top 10. On the other hand, the single machine, multi-class, approach reach 30% of the positions in the top 10. Finally, the class binarization method one-versus-all only represents the 10% but ranks first in the three MCDM methods. Note that these percentages differ slightly in GRA (one-versus-one represents 50% and one-versus-all 20%). Regarding the classifiers, the MLP, in its different configurations, reach 80% of the positions in the top 10 (70% in the case GRA). Finally, 40% of the classification models use feature selection (30%

in the case of GRA). These results demonstrate the effectiveness of the methods proposed in this research.

Table E.2: TOP 10 alternatives ranked by TOPSIS. Decoding methods in one-versus-one class binarization are labeled as: ^HHamming, ^LLoss-based and ^TThreshold.

	Value	Method	Classifier	Filter	Acc	TPR	TNR	Prec	F	AUC	Time
1	0.9966	1vsAll	MLP	None	0.96	0.92	0.97	0.92	0.93	0.95	118.12
2	0.9723	Single	MLP	None	0.96	0.91	0.96	0.90	0.92	0.94	125.83
3	0.9676	1vs1 ^H	SVM	None	0.96	0.91	0.96	0.90	0.93	0.95	773.42
4	0.9663	Single	MLP	CFS	0.95	0.91	0.96	0.90	0.91	0.94	116.18
5	0.9570	1vs1 ^L	MLP	CFS	0.95	0.90	0.97	0.91	0.90	0.93	223.71
6	0.9546	1vs1 ^H	MLP	None	0.95	0.91	0.97	0.89	0.90	0.94	221.36
7	0.9543	1vs1 ^L	MLP	None	0.95	0.91	0.97	0.89	0.90	0.94	298.69
8	0.9510	Single	Fisher	None	0.95	0.88	0.96	0.90	0.92	0.94	6.70
9	0.9374	1vs1 ^H	MLP	CFS	0.95	0.89	0.96	0.88	0.90	0.93	187.08
10	0.9363	1vs1 ^L	MLP	CFS	0.94	0.89	0.96	0.90	0.89	0.92	185.72

Table E.3: TOP 10 alternatives ranked by GRA. Decoding methods in one-versus-one class binarization are labeled as: ^HHamming, ^LLoss-based and ^TThreshold.

	Value	Method	Classifier	Filter	Acc	TPR	TNR	Prec	F	AUC	Time
1	0.9998	1vsAll	MLP	None	0.96	0.92	0.97	0.92	0.93	0.95	118.12
2	0.9668	1vs1 ^H	SVM	None	0.96	0.91	0.96	0.90	0.93	0.95	773.42
3	0.9515	Single	MLP	None	0.96	0.91	0.96	0.90	0.92	0.94	125.83
4	0.9328	1vs1 ^H	MLP	None	0.95	0.91	0.97	0.89	0.90	0.94	221.36
5	0.9328	1vs1 ^L	MLP	None	0.95	0.91	0.97	0.89	0.90	0.94	298.69
6	0.9327	1vs1 ^L	MLP	CFS	0.95	0.90	0.97	0.91	0.90	0.93	223.71
7	0.9324	Single	MLP	CFS	0.95	0.91	0.96	0.90	0.91	0.94	116.18
8	0.9243	Single	Fisher	None	0.95	0.88	0.96	0.90	0.92	0.94	6.70
9	0.8963	1vs1 ^H	MLP	CFS	0.95	0.89	0.96	0.88	0.90	0.93	187.08
10	0.8868	1vsAll	Fisher	None	0.95	0.84	0.96	0.90	0.91	0.92	16.62

Table E.5 summarizes the top 10 ranks obtained by the decision-making methods used in this research. As can be seen, they agree on the winner but the global agreement is only 20%. Therefore, the next step uses Spearman's rank correlation coefficient to generate weighted ranking in an attempt to resolve the disagreements.

E.3.2 Conflict handling results

The goal of Spearman's rank correlation coefficient is to determine the optimal weight for every MCDM method. Before the computation, the ranking scores of TOPSIS and GRA are normalized using $\frac{x-\min}{\max-\min}$ and VIKOR using $\frac{\max-x}{\max-\min}$. The weights of every MCDM method is based on the normalized ranking scores. Table E.6 shows the weights and normalized weights of every MCDM method. Note that

Table E.4: TOP 10 alternatives ranked by VIKOR. Decoding methods in one-versus-one class binarization are labeled as: ^HHamming, ^LLoss-based and ^TThreshold.

Rank	Value	Method	Classifier	Filter	Acc	TPR	TNR	Prec	F	AUC	Time
1	0.0000	1vsAll	MLP	None	0.96	0.92	0.97	0.92	0.93	0.95	118.12
2	0.0332	1vs1 ^H	SVM	None	0.96	0.91	0.96	0.90	0.93	0.95	773.42
3	0.0371	Single	MLP	None	0.96	0.91	0.96	0.90	0.92	0.94	125.83
4	0.0423	Single	MLP	CFS	0.95	0.91	0.96	0.90	0.91	0.94	116.18
5	0.0494	1vs1 ^L	MLP	CFS	0.95	0.90	0.97	0.91	0.90	0.93	223.71
6	0.0532	Single	Fisher	None	0.95	0.88	0.96	0.90	0.92	0.94	6.70
7	0.0549	1vs1 ^H	MLP	None	0.95	0.91	0.97	0.89	0.90	0.94	221.36
8	0.0549	1vs1 ^L	MLP	None	0.95	0.91	0.97	0.89	0.90	0.94	298.69
9	0.0778	1vs1 ^H	MLP	CFS	0.95	0.89	0.96	0.88	0.90	0.93	187.08
10	0.0797	1vs1 ^L	MLP	CFS	0.94	0.89	0.96	0.90	0.89	0.92	185.72

Table E.5: MCDM rankings and values.

Method	Classifier	Filter	TOPSIS		GRA		VIKOR	
			Rank	Value	Rank	Value	Rank	Value
1vsAll	MLP	None	1	0.9966	1	0.9998	1	0.0000
Single	MLP	None	2	0.9723	3	0.9515	3	0.0371
1vs1 ^H	SVM	None	3	0.9676	2	0.9668	2	0.0332
Single	MLP	CFS	4	0.9663	7	0.9324	4	0.0423
1vs1 ^L	MLP	CFS	5	0.9570	6	0.9327	5	0.0494
1vs1 ^H	MLP	None	6	0.9546	4	0.9328	7	0.0549
1vs1 ^L	MLP	None	7	0.9543	5	0.9328	8	0.0549
Single	Fisher	None	8	0.9510	8	0.9243	6	0.0532
1vs1 ^H	MLP	CFS	9	0.9374	9	0.8963	9	0.0778
1vs1 ^L	MLP	CFS	10	0.9363	11	0.8846	10	0.0797

the normalized weights are quite similar but, in light of the values of the MCDM methods, small variations may have a large impact in the ranking. Each weighted MCDM method is then applied to re-rank the alternatives using as inputs the previous rank values generated by the MCDM (see Tables E.2, E.3 and E.4).

Table E.6: Weights and normalized weights of every MCDM method.

	TOPSIS	GRA	VIKOR
Weights	0.9896	0.9910	0.9853
Normalized weights	0.3337	0.3341	0.3322

Table E.7 shows the top 10 alternatives re-ranked by weighted TOPSIS, GRA and VIKOR. As can be seen, the three rankings now agree. The level of disagreement on the rankings is dramatically reduced when using weighted MCDM methods via

Spearman’s rank correlation coefficient. Specifically, the global agreement on the top 10 alternatives of the three MCDM methods have changed from 2 to 10.

Table E.7: Weighted MCDM rankings and values.

Method	Classifier	Filter	TOPSIS		GRA		VIKOR	
			Rank	Value	Rank	Value	Rank	Value
1vsAll	MLP	None	1	1.0000	1	1.0000	1	0.0000
1vs1 ^H	SVM	None	2	0.9652	2	0.9303	2	0.0436
Single	MLP	None	3	0.9594	3	0.9187	3	0.0588
Single	MLP	CFS	4	0.9496	4	0.8981	4	0.0798
1vs1 ^L	MLP	CFS	5	0.9441	5	0.8888	5	0.0823
1vs1 ^H	MLP	None	6	0.9404	6	0.8845	6	0.0835
1vs1 ^L	MLP	None	7	0.9403	7	0.8842	7	0.0837
Single	Fisher	None	8	0.9384	8	0.8775	8	0.0924
1vs1 ^H	MLP	CFS	9	0.9133	9	0.8396	9	0.1269
1vs1 ^L	MLP	CFS	10	0.9086	10	0.8314	10	0.1391

E.4 Conclusions

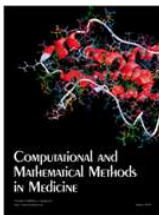
A methodology for evaluating classification problems has been presented. Its effectiveness has been demonstrated in tear film lipid layer classification. For this problem, four binarization techniques, three feature selection filters, and five machine learning algorithms have been used. Their performance was analyzed on several measures: accuracy, TPR, TNR, precision, F-measure, AUC and training time. Since this analysis involves more than one criterion, three MCDM methods were used. When the MCDM methods produced different rankings, the Spearman’s rank correlation coefficient was used to resolve disagreements.

Results showed that class binarization and feature selection play an important role in improving the performance of machine learning classifiers in tear film lipid layer classification. In particular, class binarization improves the classification criteria at the expense of a longer training time. On the other hand, feature selection dramatically reduces the training time at the expense of a slight degradation in classification performance. Note however that in some cases the performance on these criteria are maintained. Finally, the use of class binarization along with feature selection obtains a good trade-off between classification performance and training time. The MCDM methods have demonstrated to be powerful tools for combining multiple criteria. Moreover, Spearman’s rank correlation coefficient is able to improve the agreement among different rankings and provide a single answer.

Appendix F

Publications and other mentions

JCR Journals



B. Remeseiro, M. Penas, A. Mosquera, J. Novo, M. G. Penedo, E. Yebra-Pimentel. Statistical comparison of classifiers applied to the interferential tear film lipid layer automatic classification. *Computational and Mathematical Methods in Medicine*, 2012, art. 207315, 1-10, 2012.

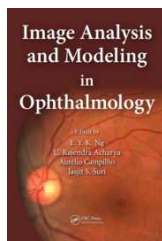


B. Remeseiro, M. Penas, N. Barreira, A. Mosquera, J. Novo, C. García-Resúa. Automatic classification of the interferential tear film lipid layer using colour texture analysis. *Computer Methods and Programs in Biomedicine*, 111, 93-103, 2013.

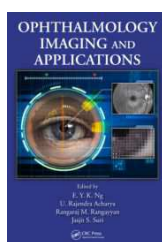


B. Remeseiro, V. Bolón-Canedo, D. Peteiro-Barral, A. Alonso-Betanzos, B. Guijarro-Berdiñas, A. Mosquera, M. G. Penedo, N. Sánchez-Marño. A Methodology for Improving Tear Film Lipid Layer Classification. *IEEE Journal of Biomedical and Health Informatics* (in press)

Book chapters



M. G. Penedo, B. Remeseiro, L. Ramos, N. Barreira, C. García-Resúa, E. Yebra-Pimentel, A. Mosquera. Automatization of Dry Eye Syndrome Tests. *Image Analysis and Modeling in Ophthalmology*, Chapter 16, 293-320, 2014.

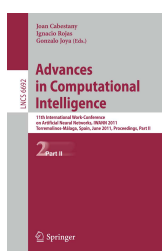


B. Remeseiro, M. G. Penedo, C. García-Resúa, E. Yebra-Pimentel, A. Mosquera. Dry Eye Characterisation by Analysing Tear Film Images. *Ophthalmology Imaging and Applications*, Chapter 23 (in press).

Chapters in Book Series



D. Calvo, A. Mosquera, M. Penas, C. García-Resúa, B. Remeseiro. Color texture analysis for tear film classification: a preliminary study. *Lecture Notes in Computer Science: International Conference on Image Analysis and Recognition (ICIAR)*, 6112, 388-397, 2010.



L. Ramos, M. Penas, B. Remeseiro, A. Mosquera, N. Barreira, E. Yebra-Pimentel. Texture and color analysis for the automatic classification of the eye lipid layer. *Lecture Notes in Computer Science: Advances in Computational Intelligence (International Work Conference on Artificial Neural Networks-IWANN 2011)*, 6692, 66-73, 2011.



B. Remeseiro, L. Ramos, N. Barreira, A. Mosquera, E. Yebra-Pimentel. Colour Texture Segmentation of Tear Film Lipid Layer Images. *Lecture Notes in Computer Science: Computer Aided Systems Theory, Revised Selected Papers EUROCAST 2013* (in press).



Rebeca Méndez, B. Remeseiro, D. Peteiro-Barral, M. G. Penedo. Evaluation of class binarization and feature selection in tear film classification using TOPSIS. Communications in Computer and Information Science: Agents and Artificial Intelligence, ICAART 2013, Revised Selected Papers (in press).

International conferences



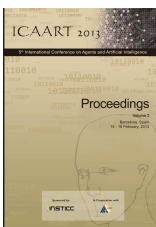
B. Remeseiro, L. Ramos, M. Penas, E. Martínez, M. G. Penedo, A. Mosquera. Colour texture analysis for classifying the tear film lipid layer: a comparative study. International Conference on Digital Image Computing: Techniques and Applications (DICTA 2011), 268-273, Noosa, Australia, December 2011.



V. Bolón-Canedo, D. Peteiro-Barral, B. Remeseiro, A. Alonso-Betanzos, B. Guijarro-Berdiñas, A. Mosquera, M. G. Penedo, N. Sánchez-Marroño, "Interferential Tear Film Lipid Layer Classification: an Automatic Dry Eye Test", IEEE 24th International Conference on Tools with Artificial Intelligence (ICTAI 2012), 359-366, Athens, Greece, November 2012.



B. Remeseiro, L. Ramos, N. Barreira, A. Mosquera, E. Yebra-Pimentel. Colour Texture Segmentation of Tear Film Lipid Layer Images. Eurocast 2013, 260-261, Las Palmas, Spain, February 2013.



R. Méndez, B. Remeseiro, D. Peteiro-Barral, M. G. Penedo. Multi-criteria evaluation of class binarization and feature selection in tear film lipid layer classification. 5th International Conference on Agents and Artificial Intelligence (ICAART 2013), 2, 62-70, Barcelona, Spain, February 2013.



B. Remeseiro, M. G. Penedo, C. García-Resúa, E. Yebra-Pimentel. An Automatic Dry Eye Test based on Lipid Layer Pattern Assessment. International Conference of Optometry and Vision Science (CIOCV 2013), 60, Braga, Portugal , April 2013.



B. Remeseiro, M. G. Penedo, C. García-Resúa, E. Yebra-Pimentel. iDEAS: Dry Eye Assessment System. VII Iberoamerican Conference on Optics - XI Latinamerican meeting on Optics, Lasers and Applications (RIO-OPTILAS 2013), Porto, Portugal, July 2013.



V. Bolón-Canedo, B. Remeseiro, N. Sánchez-Marño, A. Alonso-Betanzos. mC-ReliefF: An Extension of ReliefF for Cost-Based Feature Selection. 6th International Conference on Agents and Artificial Intelligence (ICAART 2014), 42-51, Angers, France, March 2014.



B. Remeseiro, A. Mosquera, M. G. Penedo, C. García-Resúa. Tear film maps based on the lipid interference patterns. 6th International Conference on Agents and Artificial Intelligence (ICAART 2014), 732-739, Angers, France, March 2014.



B. Remeseiro, M. G. Penedo, C. García-Resúa, E. Yebra-Pimentel. An automated tool for tear film distribution maps. II International Conference on Applications of Optics and Photonics (AOP 2014), Aveiro, Portugal, May 2014 (in press).

National conferences



B. Remeseiro, L. Ramos, N. Barreira, A. Mosquera, M. G. Penedo. Automatización de tests para el diagnóstico del síndrome de ojo seco. BioIntegraSaúde (BIS 2013), 49, Santiago de Compostela, Spain, April 2013.

JCR Journals (under review process)

D. Peteiro-Barral, B. Remeseiro, M. G. Penedo, R. Méndez. Evaluation of Tear Film Lipid Layer Classification using MCDM Methods and Rank Correlation. *Expert Systems with Applications*, 2014.

B. Remeseiro, A. Mosquera, M. G. Penedo. Seeded region growing for tear film local maps. *IEEE Transactions on Medical Imaging*, 2014.

Software registration

Software registration of the product *Herramienta para la clasificación automática de la película lagrimal*. Authorship shared by Universidade da Coruña, Manuel Francisco González Penedo, Marcos Ortega Hortas, Noelia Barreira Rodríguez and Beatriz Remeseiro López.

Appendix G

Resumen

Los ojos son indudablemente uno de los más delicados, sensibles y complejos órganos que poseemos. Son como la ventana a través de la cual vemos el mundo, y son responsables de cuatro quintos de toda la información que nuestro cerebro recibe. Por esta razón, probablemente confiamos más en nuestra vista que en cualquier otro sentido. La superficie del ojo, denominada superficie ocular, está formada por la cornea y la conjuntiva. Es un componente extraordinario y vital de la visión. Como mucosa, está protegida por el sistema inmune que usa mecanismos innatos y adaptables presentes en la película lagrimal.

Las lágrimas se segregan de la glándula lagrimal y se distribuyen mediante el parpápedo para formar la película lagrimal de la superficie ocular. La película lagrimal es responsable de mantener húmeda la superficie ocular, que es la primera línea de defensa, y también es esencial para una clara visión. Su capa más externa, denominada capa lipídica de la película lagrimal, está compuesta por una fase polar con propiedades humectantes y cubierta por una fase no polar. Se trata de la capa más fina de la película lagrimal y está principalmente cubierta por las glándulas de Meibomio, embebidas en los platos tarsales superiores e inferiores.

Un cambio cuantitativo o cualitativo en la capa lipídica normal tiene un efecto negativo en la calidad de la visión medido como sensibilidad de contraste, y en la evaporación de las lágrimas de la superficie ocular. En efecto, se ha demostrado que una sustancial evaporación de las lágrimas causada por alteraciones de la capa lipídica es característico del ojo seco evaporativo. Esta enfermedad supone una irritación de la superficie ocular, y está asociada con síntomas de malestar y sequedad. Es una dolencia común entre los adultos de mediana edad y de edades más avanzadas, y afecta a un amplio rango de la población: entre un 10% y un 20% de la población, aunque en poblaciones asiáticas este porcentaje puede alcanzar el 33%. Afecta especialmente a los usuarios de lentes de contacto, y empeora con la edad.

Las condiciones actuales de trabajo, como el uso de ordenadores, han incrementado la proporción de gente afectada de ojo seco evaporativo.

G.1 Aspectos clínicos

El síndrome de ojo seco es una enfermedad multifactorial, por lo que se necesitan varias pruebas clínicas para obtener un diagnóstico. Existe un amplio rango de pruebas que evalúan diferentes aspectos de la película lagrimal, los cuales pueden ser agrupados en dos categorías, dependiendo de los parámetros de la película lagrimal que se midan. Por una parte, las pruebas cuantitativas de la película lagrimal están relacionadas con la función de secreción de la glándula lagrimal y miden la secreción de lágrimas de la película lagrimal. Por otra parte, las pruebas cualitativas miden la habilidad de la película lagrimal de permanecer estable, lo cual es esencial para cubrir la parte anterior del ojo y realizar sus funciones.

Dentro de las pruebas clínicas cabe destacar la denominada evaluación de los patrones de la capa lipídica, que permite evaluar la calidad de la película lagrimal y el grosor de la capa lipídica mediante la observación no invasiva de la superficie de la capa lipídica por interferometría. Diferentes aparatos, basados en principios ópticos, han sido diseñados para evaluar los patrones de la capa lipídica mediante el fenómeno de interferencia. El Tearscope Plus es el instrumento utilizado por el equipo de la Facultad de Óptica y Optometría (Universidad de Santiago de Compostela) que ha colaborado en esta investigación.

El Tearscope Plus fue diseñado por Guillon como un instrumento multiuso para la examinación no invasiva de la película lagrimal, su apariencia, volumen, estabilidad, y su efecto en la superficie ocular y las lentes de contacto. Es un instrumento portátil que se puede utilizar solo o junto con un biomicroscopio. Proyecta un fuente cilíndrica de luz blanca fluorescente en la capa lipídica. El fenómeno de interferencia observado es único debido a la fuente de luz específica de este aparato.

Los optometristas necesitan reconocer distintos tipos de patrones de interferencia observables con instrumentos como el Tearscope Plus: el patrón asociado a la película lagrimal más estable, que representa al mejor candidato para un uso cómodo de lentes de contacto; al patrón asociado con un incremento de evaporación y una estabilidad reducida; el patrón normal asociado con una estabilidad media; y el patrón de cobertura fina que puede no formarse de manera continua sobre una lente de contacto. Con el objetivo de facilitar esta tarea, Guillon propuso cinco categorías principales de patrones interferenciales para las evaluaciones realizadas con el Tearscope Plus. Estos patrones están basados en características morfológicas y de

color, y en orden crecimiento de grosor son: marmóreo abierto, marmóreo cerrado, fluido, amorfo y coloreado.

Aunque este método ofrece una técnica muy útil para evaluar la calidad y la estructura de la película lagrimal, está afectada por la interpretación subjetiva del observador. Las capas lipídicas más gruesas son fácilmente observables debido a que producen patrones con ondas y colores. Sin embargo, las capas más finas son difíciles de visualizar, debido a que las franjas de color y otras características morfológicas no están presentes.

G.2 Tesis

El grosor de la capa lipídica se puede evaluar mediante la clasificación de los patrones interferenciales en una de las cinco categorías definidas por Guillon. Sin embargo, la clasificación en una de esas categorías es una tarea clínica difícil, especialmente con las capas más finas que carecen de características de color y/o morfológicas. La interpretación subjetiva de los expertos, a través de una evaluación visual, puede afectar el resultado de la clasificación. Esta tarea que consume mucho tiempo es muy dependiente del entrenamiento y de la experiencia de los optometristas, y por tanto produce un alto grado de inter- e intra- variabilidad entre observadores. El desarrollo de un método sistemático y objetivo para análisis y clasificación es altamente deseable, permitiendo un diagnóstico homogéneo y liberando a los expertos de esta tediosa tarea.

La propuesta de esta investigación es el diseño de un sistema automático para evaluar los patrones de la capa lipídica de la película lagrimal mediante la interpretación de las imágenes obtenidas con el Tearscope Plus. Con este objetivo, se utilizan diferentes técnicas de visión artificial, procesamiento de imagen y aprendizaje máquina para el desarrollo y la validación de las evaluaciones automáticas que se presentan a continuación.

El Capítulo 2 describe una metodología para evaluar la capa lipídica de la película lagrimal mediante la clasificación automática de las imágenes adquiridas con el Tearscope Plus en una de las categorías definidas por Guillon. El procedimiento llevado a cabo consiste en aplicar distintos modelos de color y diferentes descriptores de textura para obtener el conjunto de características representativas de cada patrón. Para la clasificación final de esas características en uno de los patrones de Guillon se propone el uso de diferentes algoritmos de aprendizaje máquina.

Esa primera aproximación proporciona buenos resultados a costa de un tiempo de procesamiento demasiado alto y de mucha memoria, debido a que hay que calcu-

lar un gran número de características de color y textura. Este hecho hace que la metodología inicialmente propuesta no pueda ser utilizada en aplicaciones prácticas, impidiendo su uso en rutinas clínicas. Por este motivo, la reducción de la complejidad computacional de la aproximación previa se aborda en el Capítulo 3 mediante la utilización de técnicas de reducción de la dimensión. Esta optimización se centra en la reducción de los requisitos de memoria/tiempo, de manera que no se produzca una degradación en su correcto funcionamiento.

Debido a que la heterogeneidad de la capa lipídica de la película lagrimal hace que su clasificación en una única categoría no sea siempre posible, los mapas de la película lagrimal se presentan en el Capítulo 4 con el objetivo de ilustrar la distribución local de los patrones de la capa lipídica. De esta manera, se necesitan más requisitos de memoria y tiempo a cambio de una información más detallada sobre la localización y el tamaño de los patrones interferenciales presentes en la película lagrimal.

G.3 Conclusiones

En esta tesis se han propuesto y desarrollado diferentes técnicas automáticas para la evaluación de los patrones de la capa lipídica de la película lagrimal. Estas evaluaciones automáticas no pretenden invalidar la opinión de un experto en casos particulares, sino que su objetivo es servir de gran ayuda en la rutina clínica y en la investigación.

Inicialmente, se ha presentado una metodología para evaluar la capa lipídica de la película lagrimal mediante la clasificación automática de las imágenes adquiridas con el Tearscope Plus en una de las categorías de Guillon. Este proceso se lleva a cabo mediante técnicas de análisis de color y textura, y algoritmos de aprendizaje máquina. El uso de información de color mejora los resultados en comparación con el uso de imágenes en escala de grises, debido a que algunas capas contienen características de color, además de características morfológicas. Todos los métodos de análisis de textura funcionan muy bien, proporcionando resultados por encima del 90% en algunos casos. En resumen, la combinación del método de características de co-ocurrencia y el espacio de color Lab produce el mejor resultado de clasificación con una precisión máxima por encima del 96%.

Esta metodología es capaz de proporcionar resultados fiables, pero a coste de un elevado tiempo de procesado y demasiada memoria, debido a que es necesario calcular un gran número de características. Este hecho hace que la metodología no se pueda utilizar en aplicaciones prácticas y previene su uso clínico. Por este motivo,

se propone el uso de diferentes métodos de reducción de la dimensión para reducir la complejidad computacional. Esta optimización se centra en la mejora de la precisión y los requisitos de memoria/tiempo. En primer lugar, la técnica PCA se aplica como método de extracción de características. Su uso permite reducir los requisitos de memoria transformando el espacio de entrada y sin producir una degradación en el rendimiento. Sin embargo, como se aplica una transformación, el vector de características se tiene que calcular al completo y, por tanto, no hay reducción en el tiempo de procesado. Es por ello que se propone aplicar técnicas de selección de características de manera que, cuando se establece que una característica no es necesaria, se ahorra el tiempo necesario para calcularla. En concreto, se han utilizado tres populares filtros de selección de características: CFS, basado en consistencia e INTERACT. Su funcionamiento se ha evaluado mediante cinco métodos de análisis de textura y el espacio de color Lab. Los resultados así obtenidos mejoran los resultados previos en cuanto al tiempo de procesado, mientras que mantienen la precisión. Además, se ha aplicado al problema una modificación del filtro ReliefF para selección de características basada en coste, denominado mC-ReliefF. Este método permite reducir significativamente el tiempo manteniendo los buenos resultados de precisión. Cuantitativamente, el proceso *ad-hoc* de selección de características basado en el filtro CFS, que reduce el número de características de 588 a 23 sin afectar la precisión, es el que produce el mejor equilibrio entre precisión y tiempo de procesado. Permite la automatización del proceso manual con una precisión máxima superior al 97% y un tiempo de procesado inferior a 1 segundo. Por tanto, el uso de la metodología está completamente recomendado para la práctica clínica como una herramienta de ayuda al diagnóstico del síndrome de ojo seco.

Dado que la heterogeneidad de la capa lipídica de la película lagrimal hace que no siempre sea posible clasificarla en una única categoría, se presentan los mapas de la película lagrimal para ilustrar la distribución espacial de los patrones de la capa lipídica. De esta manera, se necesitan más requisitos de memoria y tiempo a cambio de una información más detallada de la localización y el tamaño de los patrones de la película lagrimal. Se proponen tres aproximaciones diferentes para abordar este problema: un sistema de votación, un sistema de votación ponderado basado en distancias y probabilidades, y una versión adaptada del algoritmo clásico de crecimiento de regiones a partir de semillas. La primera aproximación permite comprobar la viabilidad del problema, dado que proporciona mapas de la película lagrimal cualitativamente similares a las anotaciones realizadas por tres experimentados optometristas. La segunda alternativa se centra en dos variables (probabilidades y distancias), y tiene en cuenta las probabilidades multiclase proporcionadas

por un clasificador *soft*. El análisis cuantitativo llevado a cabo demuestra que el sistema produce resultados fiables con una precisión por encima del 80% en la mayoría de casos. El problema de este método es que procesa todas las ventanas dentro de la región de interés y, aunque el tiempo de obtención de las características de una ventana es casi despreciable (menos de 1 segundo), analizar todas las ventanas lleva demasiado tiempo (aproximadamente una hora en promedio). Por este motivo, se presenta una tercera alternativa basada en el algoritmo clásico de crecimiento de regiones a partir de semillas. Este algoritmo adaptado utiliza las probabilidades multiclase proporcionadas por el clasificador *soft* como criterio de homogeneidad. El método es capaz de generar mapas de distribución de la película lagrimal realmente similares a las regiones marcadas por los optometristas, con una precisión que supera el 90% en algunos casos. Además, mejora los resultados de la aproximación previa mediante una notable reducción del tiempo de procesado, que se reduce en más de un 70% (de más de 60 minutos a menos de 20). En resumen, los mapas de distribución de la película lagrimal proporcionan a los expertos una información detallada de la película lagrimal de un paciente, lo que supone una gran ayuda en el diagnóstico y tratamiento del síndrome de ojo seco.

Bibliography

Adams, R., & Bischof, L. (1994). Seeded Region Growing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(6), 641–647.

Allwein, E. L., Schapire, R. E., & Singer, Y. (2001). Reducing multiclass to binary: A unifying approach for margin classifiers. *The Journal of Machine Learning Research*, 1, 113–141.

Alpaydin, E. (2010). *Introduction to Machine Learning* (2nd ed.). The MIT Press.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society, series B*, 36, 192–236.

Biau, G. (2012). Analysis of a Random Forests Model. *Journal of Machine Learning Research*, 13, 1063–1095.

Bolon-Canedo, V., Peteiro-Barral, D., Remeseiro, B., Alonso-Betanzos, A., Guijarro-Berdiñas, B., Mosquera, A., . . . Sánchez-Marroño, N. (2012, November). Interferential Tear Film Lipid Layer Classification: an Automatic Dry Eye Test. In *IEEE International Conference on Tools with Artificial Intelligence (ICTAI)* (pp. 359–366). Athens, Greece.

Bolon-Canedo, V., Remeseiro, B., Sánchez-Marroño, N., & Alonso-Betanzos, A. (2014). mC-ReliefF: An Extension of ReliefF for Cost-Based Feature Selection. In *6th International Conference on Agents and Artificial Intelligence (ICAART)* (pp. 42–51). Angers, France.

Bolon-Canedo, V., Sánchez-Marroño, N., & Alonso-Betanzos, A. (2011). On the behavior of feature selection methods dealing with noise and relevance over synthetic scenarios. In *The 2011 International Joint Conference on Neural Networks (IJCNN)* (pp. 1530–1537).

- Borer, S., & Süsstrunk, S. (2002). *Opponent Color Space Motivated by Retinal Processing* (Vol. 1).
- Bradski, G. (2000). OpenCV. *Dr. Dobb's Journal of Software Tools*.
- Bramer, M. (2007). *Principles of Data Mining*. Springer.
- Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32.
- Brewitt, H., & Sistani, F. (2001). Dry Eye Disease: The Scale of the Problem. *Survey of Ophthalmology*, 45(2), 199–202.
- Bron, A. J. (2001). Diagnosis of Dry Eye. *Survey of Ophthalmology*, 45(2).
- Bron, A. J., & Tiffany, J. M. (2004). The contribution of meibomian disease to dry eye. *The Ocular Surface*, 2(2), 149–65.
- Bron, A. J., Tiffany, J. M., Gouveia, S. M., Yokoi, N., & Voon, L. W. (2004). Functional aspects of the tear film lipid layer. *Experimental Eye Research*, 78(3), 347–360.
- Burges, C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2), 1–47.
- Calvo, D., Mosquera, A., Penas, M., García-Resúa, C., & Remeseiro, B. (2010). Color Texture Analysis for Tear Film Classification: A Preliminary Study. In *LNCS: International Conference on Image Analysis and Recognition (ICIAR)* (Vol. 6112, pp. 388–397).
- Çesmeli, E., & Wang, D. (2001). Texture Segmentation Using Gaussian-Markov Random Fields and Neural Oscillator Networks. *IEEE Transactions on Neural Networks*, 12.
- Christen, W. G., Manson, J. E., Glynn, R. J., Ajani, U. A., Schaumberg, D. A., Sperduto, R. D., . . . Hennekens, C. H. (1998). Low-dose aspirin and risk of cataract and subtypes in a randomized trial of US physicians. *Ophthalmic Epidemiology*, 5(3), 133–142.
- Clausi, D. A., & Jernigan, M. E. (1998). A Fast Method to Determine Co-occurrence Texture Features. *IEEE Transactions on Geoscience and Remote Sensing*, 36(1), 298–300.
- Craig, J. P., & Tomlinson, A. (1997). Importance of the lipid layer in human tear film stability and evaporation. *Optometry & Vision Science*, 74, 8–13.

-
- Dash, M., & Liu, H. (2003). Consistency-based search in feature selection. *Artificial intelligence*, 151(1-2), 155–176.
- Daubechies, I. (1992). *Ten Lectures on Wavelets*. SIAM, CBMS series.
- Daugman, J. G. (1985). Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *Journal of the Optical Society of America*, 2(7), 1160–1169.
- Dietterich, T. G., & Bakiri, G. (1995a). Solving multiclass learning problems via error-correcting output codes. *Arxiv preprint cs/9501101*.
- Dietterich, T. G., & Bakiri, G. (1995b). Solving multiclass learning problems via error-correcting output codes. *Journal of Artificial Intelligence Research*, 2, 263–286.
- Feddema, J. T., Lee, C. G., & Mitchell, O. R. (1991). Weighted selection of image features for resolved rate visual feedback control. *IEEE Transactions on Robotics and Automation*, 7(1), 31–47.
- Fernandez-Caballero, J. C., Martínez, F. J., Hervás, C., & Gutiérrez, P. A. (2010). Sensitivity versus accuracy in multiclass problems using memetic pareto evolutionary neural networks. *IEEE Transactions on Neural Networks*, 21(5), 750–770.
- Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *The Journal of Machine Learning Research*, 3, 1289–1305.
- Foulks, G. N. (2007). The correlation between the tear film lipid layer and dry eye disease. *Survey of Ophthalmology*, 52(4), 369–374.
- Friedman, J. H. (1989). Regularized Discriminant Analysis. *Journal of the American Statistical Association*.
- Friedman, N., Geiger, D., & Goldszmidt, M. (1997). Bayesian network classifiers. *Machine learning*, 29, 131–163.
- Furnkranz, J. (2003). Round robin ensembles. *Intelligent Data Analysis*, 7(5), 385–403.
- Gabor, D. (1946). Theory of Communication. *Journal of Institute for Electrical Engineering*, 93, 429–457.

- García-Resúa, C., Giráldez-Fernández, M. J., Penedo, M. G., Calvo, D., Penas, M., & Yebra-Pimentel, E. (2013). New software application for clarifying tear film lipid layer patterns. *Cornea*, *32*(4), 536–546.
- García-Resúa, C., Santodomingo-Rubido, J., Lira, M., Giráldez, M. J., & Yebra-Pimentel, E. (2009). Clinical assessment of the lower tear meniscus height. *Ophthalmic and Physiological Optics*, *29*(5), 487–496.
- Gauthier, T. D. (2001). Detecting trends using spearman’s rank correlation coefficient. *Environmental Forensics*, *2*(4), 359–362.
- Gonzalez, R., & Woods, R. (2008). *Digital image processing*. Pearson/Prentice Hall.
- Goulden, C. H. (1956). *Methods of statistical analysis (2nd ed.)*. J. Wiley & Sons, Chapman & Hall.
- Guillon, J. P. (1998). Non-invasive tearscope plus routine for contact lens fitting. *Contact Lens & Anterior Eye*, *21 Suppl 1*, 31–40.
- Guyon, I., Gunn, S., Nikravesh, M., & Zadeh, L. (2006). *Feature Extraction: Foundations and Applications*. Springer Verlag.
- Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter*, *11*(1), 10–18.
- Hall, M. A. (1999). *Correlation-based feature selection for machine learning*. Unpublished doctoral dissertation, The University of Waikato.
- Haralick, R. M., Shanmugam, K., & Dinstein, I. (1973). Texture Features for Image Classification. *IEEE Transactions on Systems, Man and Cybernetics*, *3*, 610–621.
- Hering, E. (1964). *Outlines of a Theory of the Light Sense*. Harvard University Press.
- Hogg, R., & Ledolter, J. (1987). *Engineering Statistics*. MacMillan.
- Hsu, J. (1996). *Multiple Comparisons: Theory and Methods*. Chapman & Hall/CRC.

-
- Huang, C.-L., & Wang, C.-J. (2006). A GA-based feature selection and parameters optimization for support vector machines. *Expert Systems with applications*, *31*(2), 231–240.
- Hwang, C. L., & Yoon, K. (1981). *Multiple attribute decision making: methods and applications* (Vol. 13). Springer-Verlag New York.
- Inza, I., Larrañaga, P., Blanco, R., & Cerrolaza, A. J. (2004). Filter versus wrapper gene selection approaches in DNA microarray domains. *Artificial intelligence in medicine*, *31*(2), 91–103.
- Jensen, F. (1996). *An Introduction to Bayesian Networks*. Springer.
- Jie, Y., Xu, L., Wu, Y. Y., & Jonas, J. B. (2008). Prevalence of dry eye among adult Chinese in the Beijing Eye Study. *Eye*, *23*(3), 688–693.
- Jolliffe, I. T. (1986). *Principal Component Analysis*. Springer-Verlag.
- Kira, K., & Rendell, L. A. (1992). A practical approach to feature selection. In *Proceedings of the 9th International Workshop on Machine Learning* (pp. 249–256).
- Kononenko, I. (1994). Estimating attributes: analysis and extensions of RELIEF. In *Machine Learning: ECML-94* (pp. 171–182).
- Korb, D. R. (2002). *The Tear Film: Structure, Function and Clinical Examination*. Butterworth-Heinemann.
- Kotsiantis, S. B. (2007). Supervised Machine Learning: A Review of Classification Techniques. *Informatika*, *31*, 249–268.
- Kuo, Y., Yang, T., & Huang, G. W. (2008). The use of grey relational analysis in solving multiple attribute decision-making problems. *Computers & Industrial Engineering*, *55*(1), 80–93.
- Landwehr, N., Hall, M., & Frank, E. (2005). Logistic Model Trees. *Machine Learning*, *59*(1-2), 161–205.
- Larke, J. R. (1997). *The Eye in Contact Lens Wear*. Butterworth-Heinemann.
- Lemp, M. A., Baudouin, C., Baum, J., Dogru, M., Foulks, G. N., Kinoshita, S., . . . Toda, I. (2007a). The Definition and Classification of Dry Eye Disease:

- Report of the Definition and Classification Subcommittee of the International Dry Eye WorkShop (2007). *The Ocular Surface*, 5(2), 75–92.
- Lemp, M. A., Baudouin, C., Baum, J., Dogru, M., Foulks, G. N., Kinoshita, S., . . . Toda, I. (2007b). The Epidemiology of Dry Eye Disease: Report of the Epidemiology Subcommittee of the International Dry Eye WorkShop (2007). *The Ocular Surface*, 5(2), 93–107.
- Lemp, M. A., & Hamil, J. R. (1973). Factors affecting tear film breakup in normal eyes. *Archives of Ophthalmology*, 89, 103–105.
- Lilliefors, H. (1967). On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown. *Journal of the American Statistical Association*, 62(318), 399–402.
- Loughrey, J., & Cunningham, P. (2005). Overfitting in wrapper-based feature subset selection: The harder you try the worse it gets. *Research and Development in Intelligent Systems XXI*, 33–43.
- Mallat, S. G. (1989). A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11, 674–693.
- Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval* (Vol. 1). Cambridge University Press.
- Massey, F. J. (1951). The Kolmogorov-Smirnov Test for Goodness of Fit. *Journal of the American Statistical Association*, 46, 68–78.
- McDonald, J. E. (1969). Surface phenomena of the tear film. *American Journal of Ophthalmology*, 67(1), 56–64.
- McLaren, K. (1976). The development of the CIE 1976 (L*a*b) uniform colour-space and colour-difference formula. *Journal of the Society of Dyers and Colourists*, 92(9), 338–341.
- Méndez, R., Remeseiro, B., Peteiro-Barral, D., & Penedo, M. G. (2013). Multi-criteria evaluation of class binarization and feature selection in tear film lipid layer classification. In *5th International Conference on Agents and Artificial Intelligence (ICAART)* (pp. 62–70). Barcelona, Spain.
- Mengher, L. S., Bron, A. J., Tonge, S. R., & Gilbert, D. J. (1985). Effect of fluorescein instillation on the pre-corneal tear film stability. *Current Eye Research*, 4, 9–12.

-
- Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M., & Euler, T. (2006, August). YALE: Rapid Prototyping for Complex Data Mining Tasks. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD)* (pp. 935–940).
- Miller, K. (1969). *Seeing*. Oxford University Press.
- Min, F., Hu, Q., & Zhu, W. (2013). Feature selection with test cost constraint. *International Journal of Approximate Reasoning*.
- Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill.
- Moore, D. S. (1976). *Chi-Square Tests*. Defense Technical Information Center.
- Moss, S. E. (2000). Prevalence of and Risk Factors for Dry Eye Syndrome. *Archives of Ophthalmology*, 118(9), 1264–1268.
- Murthy, S. K. (1998). Automatic Construction of Decision Trees from Data: A Multi-Disciplinary Survey. *Data Mining and Knowledge Discovery*, 2, 345–389.
- Nichols, J. J., Puente, K. K. N. B., Saracino, M., & Mitchell, G. L. (2002). Evaluation of tear film interference patterns and measures of tear break-up time. *Optometry & Vision Science*, 79(6), 363–369.
- Nichols, K. K., Nichols, J. J., & Mitchell, G. L. (2004). The lack of association between signs and symptoms in patients with dry eye disease. *Cornea*, 23(8), 762–770.
- Nichols, K. K., Nichols, J. J., & Zadnik, K. (2000). Frequency of dry eye diagnostic test procedures used in various modes of ophthalmic practice. *Cornea*, 19(4), 477–482.
- Opricovic, S. (1998). Multicriteria optimization of civil engineering systems. *Faculty of Civil Engineering, Belgrade*, 2(1), 5–21.
- Opricovic, S., & Tzeng, G. H. (2004). Compromise solution by MCDM methods: A comparative analysis of VIKOR and TOPSIS. *European Journal of Operational Research*, 156(2), 445–455.
- Peng, Y., Kou, G., Wang, G., & Shi, Y. (2011). FAMCDM: A fusion approach of MCDM methods to rank multiclass classification algorithms. *Omega*, 39(6), 677–689.

- Peteiro-Barral, D., Remeseiro, B., Penedo, M. G., & Méndez, R. (n.d.). Evaluation of Tear Film Lipid Layer Classification using MCDM Methods and Rank Correlation. *Expert Systems with Applications (under review)*.
- Pflugfelder, S., Tseng, S., Sanabria, O., Kell, H., Garcia, C., Felix, C., . . . Reis, B. (1998). Evaluation of subjective assessments and objective diagnostic tests for diagnosing tear-film disorders known to cause ocular irritation. *Cornea*, *17*(1), 38–56.
- Press, W. H., Flannery, B. P., Teukolsky, S. A., & Vetterling, W. T. (1986). *Numerical recipes: the art of scientific computing*. Cambridge University Press.
- Ramos, L., Penas, M., Remeseiro, B., Mosquera, A., Barreira, N., & Yebra-Pimentel, E. (2011). Texture and Color Analysis for the automatic classification of the eye lipid layer. In *LNCS: Advances in Computational Intelligence (International Work Conference on Artificial Neural Networks, IWANN)* (Vol. 6692, pp. 66–73).
- Remeseiro, B., Bolon-Canedo, V., Peteiro-Barral, D., Alonso-Betanzos, A., Guijarro-Berdiñas, B., Mosquera, A., . . . Sánchez-Marroño, N. (2014). A Methodology for Improving Tear Film Lipid Layer Classification. *IEEE Journal of Biomedical and Health Informatics (in press)*.
- Remeseiro, B., Mosquera, A., & Penedo, M. G. (n.d.). Seeded region growing for tear film local maps. *IEEE Transactions on Medical Imaging (under review)*.
- Remeseiro, B., Mosquera, A., Penedo, M. G., & García-Resúa, C. (2014). Tear film maps based on the lipid interference patterns. In *6th International Conference on Agents and Artificial Intelligence (ICAART)* (pp. 732–739). Angers, France.
- Remeseiro, B., Penas, M., Barreira, N., Mosquera, A., Novo, J., & García-Resúa, C. (2013). Automatic classification of the interferential tear film lipid layer using colour texture analysis. *Computer Methods and Programs in Biomedicine*, *111*, 93–103.
- Remeseiro, B., Penas, M., Mosquera, A., Novo, J., Penedo, M. G., & Yebra-Pimentel, E. (2012). Statistical comparison of classifiers applied to the interferential tear film lipid layer automatic classification. *Computational and Mathematical Methods in Medicine, 2012 (art. 207315)*, 1–10.

-
- Remeseiro, B., Ramos, L., Barreira, N., Mosquera, A., & Yebra-Pimentel, E. (2013). Colour texture segmentation of tear film lipid layer images. In *Eurocast 2013* (pp. 260–261).
- Remeseiro, B., Ramos, L., Penas, M., Martínez, E., Penedo, M. G., & Mosquera, A. (2011, December). Colour texture analysis for classifying the tear film lipid layer: a comparative study. In *International Conference on Digital Image Computing: Techniques and Applications (DICTA)* (pp. 268–273). Noosa, Australia.
- Rieger, G. (1992). The importance of the precorneal tear-film for the quality of optical imaging. , *76*(3), 157–158.
- Rodriguez, J., Perez, A., & Lozano, J. (2010). Sensitivity analysis of k-fold cross-validation in prediction error estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*, 569–575.
- Rolando, M., Iester, M., Macrí, A., & Calabria, G. (1998). Low spatial-contrast sensitivity in dry eyes. , *17*(4), 376–369.
- Rolando, M., Valente, C., & Barabino, S. (2008). New test to quantify lipid layer behavior in healthy subjects and patients with keratoconjunctivitis sicca. *Cornea*, *27*(8), 866–870.
- Rosenblatt, F. (1958). The Perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*, 386–408.
- Russ, J. (1999). *The image processing handbook (3rd ed.)*. Boca Raton, FL, USA: CRC Press, Inc.
- Sangwine, S. J., & Horne, R. E. N. (1998). *The Colour Image Processing Handbook*. Chapman & Hall.
- Schaumberg, D. A., Sullivan, D. A., Buring, J. E., & Dana, M. R. (2003). Prevalence of dry eye syndrome among US women. *American Journal of Ophthalmology*, *136*, 318–326.
- Schiffman, R. M., Christianson, M. D., Jacobsen, G., Hirsch, J. D., & Reis, B. L. (2000). Reliability and validity of the Ocular Surface Disease Index. *Archives of Ophthalmology*, *118*(5), 615–621.
- Schirmer, O. (1903). Studiend zur Physiologie und Pathologie der Tranenabsonderung und Tranenabfuhr. *Graefe's Archive for Clinical and Experimental Ophthalmology*, *56*, 197–291.

Sivagaminathan, R. K., & Ramakrishnan, S. (2007). A hybrid approach for feature subset selection using neural networks and ant colony optimization. *Expert systems with applications*, 33(1), 49–60.

Smith, J. A. (2007). The Epidemiology of Dry Eye Disease: Report of the Epidemiology Subcommittee of the International Dry Eye WorkShop. *The Ocular Surface*, 5(2), 93–107.

Tearscope plus clinical handbook and tearscope plus instructions. (1997). Keeler Ltd. Windsor, Berkshire, Keeler Inc, Broomall, PA.

Teich, J. (2001). Pareto-front exploration with uncertain objectives. In *Evolutionary multi-criterion optimization* (pp. 314–328).

Tomlinson, A., Blades, K. J., & Pearce, E. I. (2001). What does the phenol red thread test actually measure? *Optometry & Vision Science*, 78(3), 142–146.

Topcon DV-3 digital video camera. (n.d.). (Topcon Medical Systems, Oakland, NJ, USA)

Topcon IMAGEnet i-base. (n.d.). (Topcon Medical Systems, Oakland, NJ, USA)

Topcon SL-D4 slit lamp. (n.d.). (Topcon Medical Systems, Oakland, NJ, USA)

VOPTICAL_I1, VARPA optical dataset annotated by optometrists from the Faculty of Optics and Optometry, University of Santiago de Compostela (Spain). (n.d.). Retrieved March 2014, from http://www.varpa.es/voptical_I1.html

VOPTICAL_Is, VARPA optical dataset annotated by optometrists from the Faculty of Optics and Optometry, University of Santiago de Compostela (Spain). (n.d.). Retrieved March 2014, from http://www.varpa.es/voptical_Is.html

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6), 80–83.

Wolff, E. (1954). *Anatomy of the eye and orbit (4th edition)*. H. K. Lewis and Co., London.

Woods, J. W. (1972). Two-dimensional discrete markovian fields. *IEEE Transactions on Information Theory*, 18(2), 232–240.

Zhao, Z., & Liu, H. (2007). Searching for interacting features. In *Proceedings of the 20th international joint conference on Artificial intelligence* (pp. 1156–1161).