José Ramon Trindade Pires

“Diabetic retinopathy image quality assessment, detection, screening and referral”

“*Análise de qualidade, detecção de lesões de retinopatia diabética, triagem e verificação de necessidade de consulta a partir de imagens de retina.*”

CAMPINAS
2013

University of Campinas
Institute of Computing

Universidade Estadual de Campinas
Instituto de Computação

José Ramon Trindade Pires

“Diabetic retinopathy image quality assessment, detection, screening and referral”

Supervisor:
Orientador(a): **Prof. Dr. Anderson Rocha**

“Análise de qualidade, detecção de lesões de retinopatia diabética, triagem e verificação de necessidade de consulta a partir de imagens de retina.”

MSc Dissertation presented to the Post Graduate Program of the Institute of Computing of the University of Campinas to obtain a Mestre degree in Computer Science.

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Computação da Universidade Estadual de Campinas para obtenção do título de Mestre em Ciência da Computação.

This volume corresponds to the final version of the Dissertation defended by José Ramon Trindade Pires, under the supervision of Prof. Dr. Anderson Rocha.

Este exemplar corresponde à versão final da Dissertação defendida por José Ramon Trindade Pires, sob orientação de Prof. Dr. Anderson Rocha.

Anderson de Rezende Rocha.

Supervisor’s signature / *Assinatura do Orientador(a)*

CAMPINAS
2013

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Matemática, Estatística e Computação Científica
Maria Fabiana Bezerra Muller - CRB 8/6162

P665d Pires, Ramon, 1989-
Diabetic retinopathy image quality assessment, detection, screening and referral / José Ramon Trindade Pires. – Campinas, SP : [s.n.], 2013.

Orientador: Anderson de Rezende Rocha.
Dissertação (mestrado) – Universidade Estadual de Campinas, Instituto de Computação.

1. Retinopatia diabética. 2. Aprendizado de máquina. 3. Visão por computador. 4. Fusão de classificadores. I. Rocha, Anderson de Rezende,1980-. II. Universidade Estadual de Campinas. Instituto de Computação. III. Título.

Informações para Biblioteca Digital

Título em outro idioma: Análise de qualidade, detecção de lesões de retinopatia diabética, triagem e verificação de necessidade de consulta a partir de imagens de retina
Palavras-chave em inglês:
Diabetic retinopathy
Machine learning
Computer vision
Classifier fusion
Área de concentração: Ciência da Computação
Titulação: Mestre em Ciência da Computação
Banca examinadora:
Anderson de Rezende Rocha [Orientador]
Hélio Pedrini
Agma Juci Machado Traina
Data de defesa: 14-06-2013
Programa de Pós-Graduação: Ciência da Computação

TERMO DE APROVAÇÃO

Dissertação Defendida e Aprovada em 14 de Junho de 2013, pela Banca examinadora composta pelos Professores Doutores:

AJM Traina

Profª. Drª. Agma Juci Machado Traina
ICMC / USP

Hélio Pedrini

Prof. Dr. Hélio Pedrini
IC / UNICAMP

Anderson de Rezende Rocha

Prof. Dr. Anderson de Rezende Rocha
IC / UNICAMP

Institute of Computing /*Instituto de Computação*
University of Campinas /*Universidade Estadual de Campinas*

Diabetic retinopathy image quality assessment, detection, screening and referral

José Ramon Trindade Pires[1]

June 14, 2013

Examiner Board/*Banca Examinadora*:

- Prof. Dr. Anderson Rocha (Supervisor/*Orientador*)
- Prof. Dr. Hélio Pedrini
  Institute of Computing - UNICAMP
- Prof.ª Dr.ª Agma Juci Machado Traina
  Institute of Mathematics and Computer Science - USP
- Prof. Dr. Ricardo da Silva Torres
  Institute of Computing - UNICAMP (Substitute/*Suplente*)
- Prof. Dr. Paulo André Vechiatto de Miranda
  Institute of Mathematics and Statistics - USP (Substitute/*Suplente*)

[1] Financial support: CAPES scholarship 2011–2012, FAPESP scholarship (process 2011/15349-3) 2012–2013

Abstract

Diabetic Retinopathy (DR), a common complication caused by diabetes, manifests through different lesions that have their particularities. These particularities are explored in the literature as methods for representation, providing a satisfactory discrimination between healthy/diseased retinas. However, by being strongly linked to the visual characteristics of each anomaly, the detection of distinct lesions requires distinct approaches. In this work, we present a general framework whose objective is to automate the eye-fundus image analysis. The work comprises four steps: image quality assessment, DR-related lesion detection, screening, and referral. In the first step, we apply characterization techniques to assess image quality by two criteria: field definition and blur detection. In the second step of this work, we extend up a previous work of our group which explored a unified method for detecting distinct lesions in eye-fundus images. In our approach for detection of any lesion, we explore several alternatives for low-level (dense and sparse extraction) and mid-level (coding/pooling techniques of bag of visual words) representations, aiming at the development of an effective set of individual DR-related lesion detectors. The scores derived from each individual DR-related lesion, taken for each image, represent a high-level description, fundamental point for the third and fourth steps. Given a dataset described in high-level (scores from the individual detectors), we propose, in the third step of the work, the use of machine learning fusion techniques aiming at the development of a multi-lesion detection method. The high-level description is also explored in the fourth step for the development of an effective method for evaluating the necessity of referral of a patient to an ophthalmologist in the interval of one year, avoiding overloading medical specialist with simple cases as well as give priority to patients in an urgent state.

Resumo

A Retinopatia Diabética (RD), complicação provocada pela diabetes, se manifesta por meio de diferentes lesões que possuem suas especificidades. Estas especificidades são exploradas na literatura como estratégia para representação, proporcionando uma discriminação satisfatória entre imagens de pacientes normais e doentes. No entanto, por estar fortemente atrelada às características visuais de cada anomalia, a detecção de lesões distintas exige abordagens distintas. Neste trabalho, apresentamos um arcabouço geral cujo objetivo é automatizar o procedimento de análise de imagens de fundo de olho. O trabalho é dividido em quatro etapas: avaliação de qualidade, detecção de lesões individuais, triagem e verificação de necessidade de consulta. Na primeira etapa, aplicamos diferentes técnicas de caracterização de imagens para avaliar a qualidade das imagens por meio de dois critérios: definição de campo e detecção de borramentos. Na segunda etapa deste trabalho, propomos a continuação de um trabalho anterior desenvolvido pelo nosso grupo, no qual foi aplicado um método unificado na tentativa de detecção de lesões distintas. No nosso método para detecção de qualquer lesão, exploramos diferentes alternativas de representação em baixo nível (extração densa e esparsa) e médio nível (técnicas de coding/pooling para sacolas de palavras visuais) objetivando o desenvolvimento de um conjunto eficaz de detectores de lesões individuais. As pontuações provenientes de cada detector de lesão, obtidas para cada imagem, representam uma descrição de alto nível, ponto fundamental para a terceira e a quarta etapas. Tendo em mãos um conjunto de dados descritos em alto nível (pontuações dos detectores individuais), propomos, na terceira etapa do trabalho, a aplicação de técnicas de fusão de dados para o desenvolvimento de um método de detecção de múltiplas lesões. A descrição em alto nível também é explorada na quarta etapa para o desenvolvimento de um método eficaz de avaliação de necessidade de encaminhamento a um oftalmologista no intervalo de um ano, visando evitar que o médico seja sobrecarregado, bem como dar prioridade a pacientes em estado urgente.

Acknowledgements

It is the end of a long and arduous stage. Restless nights, concerns about outside university issues, health problems, initial inexperience in scientific researches etc. It was a hard period of two and a half years, but at no time the word “abandonment” crossed my mind. And by the way, it has always been miles away. The reason? Friends, family, professors who always encouraged me. And here are my sincere acknowledgements.

First of all, I would like to thank God, for having always given me the strength to face the challenges, overcome them and gain experience. My God always enlightens me in the decision making.

A special thanks goes to my parents, especially to my dear mother, with whom I communicate almost every day and always asks me how my day was. *Thank, Maria Lisboa, for your prayers and advice.*

Thanks also to all my siblings. I insist on mentioning all the brothers (Ney, Nildo, Cani, Marcelo, and Valério) and sisters (Gracinha, Têca, Mônica, Livinha, Ana, and Débora), of whom I am very proud for our union and for the happiness they provide me. *It is good to know I can always count on you.* Thanks also to my grandmother Aurelina and my nieces and nephews (yes, there are many).

I would like to thank my advisor Anderson for having motivated me to come to Campinas, the Brazilian Silicon Valley, when I was still in doubt about which university I would do the master's course at. *Thank you, Anderson, for the patience and great help.*

I could not forget here the RECOD colleagues, Peruvian, Bolivian, Costa Rican, Colombian, and Brazilian friends of the Institute of Computing. Thanks also to the professors Jacques Wainer, Eduardo Valle and Siome Goldenstein for the cooperation. Thanks a lot also to Herbert Jelinek, who helped me many times and told me he is looking forward my possible visit to Australia during my doctoral studies.

Thanks also to the professors of the State University of Southwest Bahia, specially to Roque Trindade, who told me and some of my classmates: *“You are the elite”.*

I can not forget to mention the gratitude to Unicamp. This is a university that supports the student at all times. I would also like to thank FAPESP (grant number 2011/15349-3) and CAPES for financial support.

Contents

Abstract **xi**

Resumo **xiii**

Acknowledgements **xv**

1 Introduction **1**

1.1 DR-related Lesions 3

1.2 Stages of the Work 6

1.2.1 Quality Assessment 6

1.2.2 Quality Improvement 7

1.2.3 DR-related Lesion Detectors 7

1.2.4 Detector Fusion 7

1.2.5 Referral 8

1.3 Overview 8

2 Retinal Image Quality Analysis for Automatic Diabetic Retinopathy Detection **9**

2.1 Preamble 9

2.2 Related work 10

2.3 Technique for Field Definition 11

2.3.1 Characterization 12

2.3.2 Learning 13

2.4 Technique for Blur Detection 13

2.4.1 Characterization 13

2.4.2 Learning 17

2.4.3 Fusion 17

2.5 Experiments and validation 18

2.5.1 Datasets 18

2.5.2 Round #1: Single Results for Field Definition 19

2.5.3 Round #2: Cross-dataset Results for Field Definition 20
2.5.4 Round #3: Single Results for Blur Detection 22
2.5.5 Round #4: Cross-dataset Results for Blur Detection 24
2.6 Final Remarks 26

3 Quality Control and Multi-lesion Detection in Automated Retinopathy Classification 29
3.1 Preamble 29
3.2 Proposed Methodology 30
3.2.1 Quality selection 30
3.3 Results 31
3.3.1 Dataset 31
3.3.2 Image Quality 32
3.4 Final Remarks 33

4 Advancing Bag-of-Visual-Words Representations for Lesion Classification in Retinal Images 35
4.1 Preamble 35
4.2 Related work 37
4.2.1 Diabetic Retinopathy 37
4.2.2 BoVW Representations 40
4.3 Proposed Methodology 42
4.3.1 BoVW-based representation 42
4.3.2 Semi-soft coding 44
4.3.3 Class-aware codebook 45
4.4 Experiments 47
4.4.1 Data, protocol and metrics 47
4.4.2 Results 49
4.5 Final Remarks 51

5 Data Fusion for Multi-lesion Diabetic Retinopathy Detection 53
5.1 Preamble 53
5.2 Related work 54
5.3 Proposed Methodology 55
5.3.1 Fusion of detectors 55
5.4 Experimental results 56
5.5 Final Remarks 60

6 Assessing the Need for Referral on Diabetic Retinopathy Treatment 61

6.1 Preamble 61

6.2 Related work 62

6.3 Proposed Methodology 64

6.3.1 Detection of Individual DR-related Lesions 64

6.3.2 High-level Feature Extraction 65

6.3.3 Normalization 66

6.3.4 Meta-classification 67

6.4 Validation and Experiments 67

6.4.1 Datasets 67

6.4.2 5 × 2-Folds Cross-Validation 67

6.4.3 Experiments 69

6.5 Final Remarks 74

7 Conclusions 81

Bibliography 84

List of Tables

2.1 Field definition: AUC for the experiments. 21
2.2 Blur Detection: AUC for the experiments. 25

3.1 Blur Detection: AUC for more recent experiments. 33

4.1 Annotation occurrences for the three datasets 48
4.2 Composition of the cross-dataset training and test 48
4.3 AUCs in %, for Training with DR1, Testing with DR2 50
4.4 AUCs in %, for Training with DR1, Testing with Messidor 50

6.1 Annotation occurrences for DR2 dataset 69
6.2 AUCs for individual detectors 70
6.3 AUCs for referral 72

List of Figures

1.1 Regions and global projections of the number of people with diabetes by 2030. 2
1.2 Example of a healthy retina with its typical anatomical elements. 3
1.3 Example of a retinal image with Hard Exudates. 4
1.4 Example of a retinal image with Hemorrhages. 4
1.5 Example of a retinal image with Cotton Wool Spots. 5
1.6 Example of a retinal image with Drusen. 5
1.7 Example of a retinal image with Neovascularization. 6
1.8 Overview of the work. 8

2.1 Examples of RoIs whose images are centered on the macula (left), centered on the optic disc (middle), and non-representative (right). 12
2.2 Retina with enough quality (left) and with blurring (right) with their respective Canny edge maps (inverted for visualization purposes). 14
2.3 DR1 field definition using 5-fold cross-validation. 20
2.4 DR2 field definition using 5-fold cross-validation. 21
2.5 Cross-dataset validation for field definition using DR1 as training and DR2 as testing sets. 22
2.6 DR1 blur classification using 5-fold cross-validation. 23
2.7 DR2 blur classification using 5-fold cross-validation 24
2.8 Cross-dataset validation for blur classification using DR1 as training and DR2 as testing sets. 26

3.1 Input image with its Canny edge map as well as the 50 50×50-selected image regions centered on the edges (small squared regions) and the calculated SURF PoIs within each region (green circles), followed by a highlighted patch. 31
3.2 Image quality analysis considering 50, 50×50-regions per image from DR2 data set and various dictionary (most representative regions) sizes. 32

4.1 Scheme of previous work of our group for the detection of bright and red lesion, exploiting only the hard-sum assignment technique [72]. 36

4.2 The BoVW model illustrated in a convenient matrix form, highlighting the relationships between the low-level features $\mathbf{x}_j$, the codewords $\mathbf{c}_m$ of the visual dictionary, the encoded features α_m, the coding function f and the pooling function g. 41

4.3 Regions of interest marked by a medical specialist (dashed black regions) and the points of interest extracted in the sparse technique (blue circles). Points of interest falling within the regions marked by the specialist are further considered for creating the BoVW representation of a lesion while points found in normal images are used for the BoVW representation of images of healthy patients. In the class-aware codebook, both representation are combined. 46

4.4 Box-plot for the within-subjects (per lesion) standardized AUCs for six combinations of feature extraction and coding, with averages and 95%-confidence intervals (in red), the whiskers show the range up to 1.5× the interquartile range, and outliers are shown as small circles. The strong synergy between sparse feature extraction and semi-soft coding is evident. This plot is based on a balanced design with the DR2 dataset and all lesions, another balanced design with both datasets and two lesions show similar results. 51

5.1 Example of an SVM classifier with linear kernel (lesion vs. normal). 57

5.2 Results considering the logic OR fusion for the considered anomalies in three different scenarios. 58

5.3 Results for the meta-classifier trained with the outcome of individual DR-based lesion detectors for DR2 dataset. 59

6.1 An overview of the referable vs. non-referable approach. 68

6.2 ROC assessing the selectivity and sensitivity for the final referral decision, when using hard-sum mid-level BoVW feature vectors, and no normalization for the high-level feature vector of scores. Cross-validation with 5×2-folds. 71

6.3 ROC assessing the selectivity and sensitivity for the final referral decision, when using soft-max mid-level BoVW feature vectors, and no normalization for the high-level feature vector of scores. Cross-validation with 5×2-folds. 72

6.4 ROC assessing the selectivity and sensitivity for the final referral decision, when using hard-sum mid-level BoVW feature vectors, and term-frequency normalization for the high-level feature vector of scores. Cross-validation with 5×2-folds. 73

6.5 ROC assessing the selectivity and sensitivity for the final referral decision, when using soft-max mid-level BoVW feature vectors, and term-frequency normalization for the high-level feature vector of scores. Cross-validation with 5 × 2-folds. 74

6.6 ROC assessing the selectivity and sensitivity for the final referral decision, when using hard-sum mid-level BoVW feature vectors, and z-scores (standard normalization) for the high-level feature vector of scores. Cross-validation with 5 × 2-folds. 75

6.7 ROC assessing the selectivity and sensitivity for the final referral decision, when using soft-max mid-level BoVW feature vectors, and z-scores (standard normalization) for the high-level feature vector of scores. Cross-validation with 5 × 2-folds. 76

6.8 Arithmetic mean and standard deviation of the ROC curves for the final referral decision, when using hard-sum mid-level BoVW features considering term-frequency and z-score normalization techniques as well as no-normalization. 77

6.9 Arithmetic mean and standard deviation of the ROC curves for the final referral decision, when using soft-max mid-level BoVW features considering term-frequency and z-score normalization techniques as well as no-normalization. 78

6.10 ROC curve resulting from [22], obtained on a specific dataset (different from ours). The blue star shows the result obtained by a human expert on a subset extracted randomly from the same dataset. Note that this human expert was not exactly in the same conditions of the interpretation of the system. 79

Chapter 1

Introduction

Diabetes mellitus (DM) is a chronic end-organ disease caused by a decrease in insulin sensitivity or a loss of pancreatic function, depending on the type of diabetes, both leading to an increase in the blood glucose level. An increased blood sugar level may lead to damage of blood vessels in all organ systems of the body. The disease has thus attracted the interest of both the Health-care and Engineering communities.

Currently, diabetes affects 366 million people worldwide or 8.3% of adults. It is estimated that this number will increase to approximately 552 million people (one adult in 10 worldwide will have diabetes), according to the International Diabetes Federation (IDF)[1]. The largest increases will take place in the regions dominated by developing economies. Fig. 1.1 depicts projections of the number of people with diabetes[2] for each region by 2030.

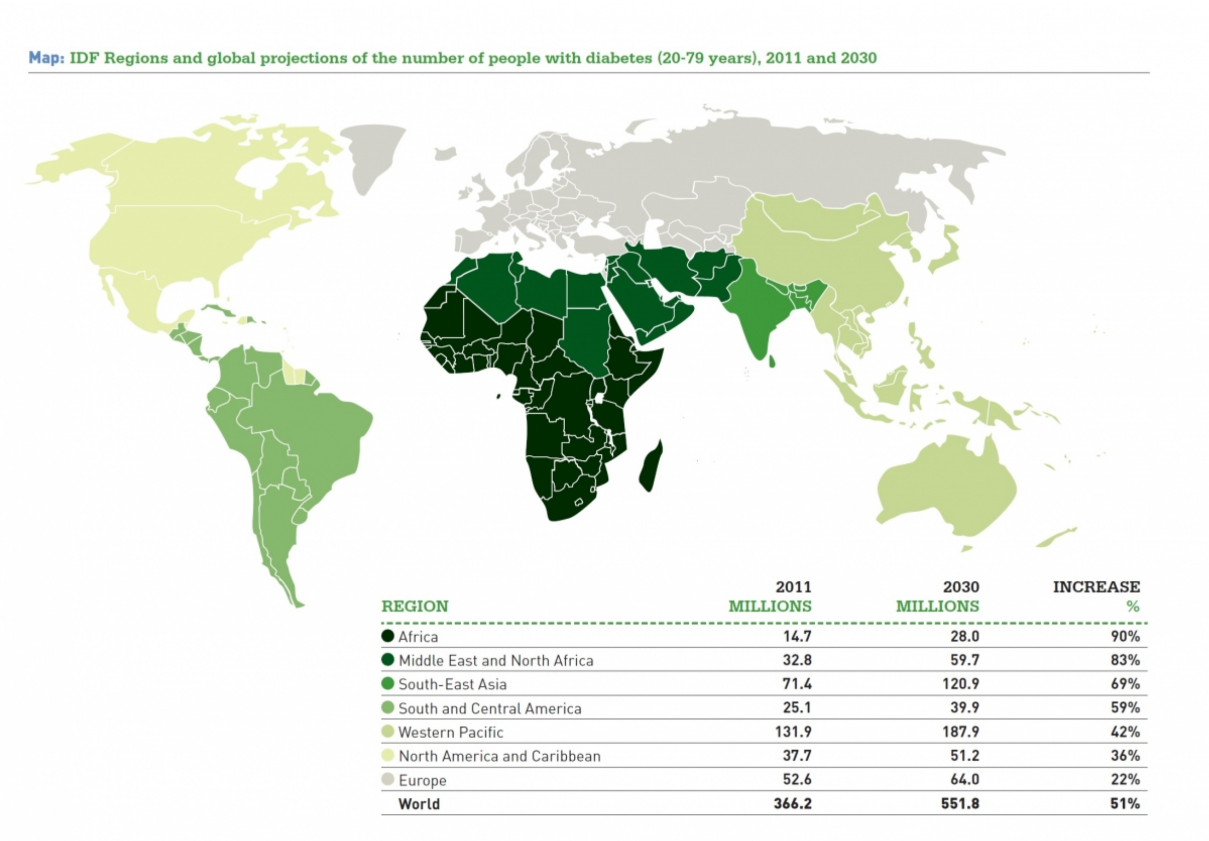
The World Health Organization (WHO)[3] projects that diabetes deaths will double between 2005 and 2030 [91].

The growing prevalence of diabetes creates an increasing prevalence of the complications related to the disease, including Diabetic Retinopathy (DR). DR occurs in approximately 2-4% of the population but is greater in indigenous populations according to some studies [94, 82]. Recent reports have shown that, in the United States, approximately 25,000 people with diabetes go blind every year due to DR [1]. Furthermore, also in the United States, the number of 40-year or older Americans with DR is projected to triple from 5.5 million in 2005 to 16 million by 2050 [74]. DR is the main cause of blindness in the 20 to 74 age group in developed countries, creating the need for systems that screen diabetic retinopathy in its early stages, so to allow an economically viable management of the disease [61].

[1] `http://www.idf.org/diabetesatlas/5e/diabetes`

[2] Figure extracted from IDF website

[3] `http://www.who.int/diabetes/en/index.html`



Map: IDF Regions and global projections of the number of people with diabetes (20-79 years), 2011 and 2030

REGION	2011 MILLIONS	2030 MILLIONS	INCREASE %
● Africa	14.7	28.0	90%
● Middle East and North Africa	32.8	59.7	83%
● South-East Asia	71.4	120.9	69%
● South and Central America	25.1	39.9	59%
● Western Pacific	131.9	187.9	42%
● North America and Caribbean	37.7	51.2	36%
● Europe	52.6	64.0	22%
World	366.2	551.8	51%

Figure 1.1: Regions and global projections of the number of people with diabetes by 2030.

It is estimated that in 2002, diabetic retinopathy accounted for about 5% of world blindness, representing almost 5 million people blind. Nowadays, according to the Diabetic Programs of the World Health Organization, DR is a leading cause of blindness, amputation and kidney failure.

According to the U.S. National Eye Institute (NEI)[4], the DR has four stages:

- *Mild Nonproliferative Retinopathy:* This corresponds to the earliest stage of the disease, in which the microaneurysms (small areas of balloon-like swelling) occurs.

- *Moderate Nonproliferative Retinopathy:* Second stage of the disease in which the blood vessels responsible to nourish the retina are blocked.

- *Severe Nonproliferative Retinopathy:* The third stage of the disease in which there is the blocking of many more blood vessels, depriving several areas of the retina

[4] `http://www.nei.nih.gov/health/diabetic/retinopathy.asp`

with their blood supply. Because of this poor irrigation, some areas of the retina send signals to the body to make growing new blood vessels.

- *Proliferative Retinopathy:* At this advanced stage, the signals sent by the retina for nourishment trigger the growth of new blood vessels. These new blood vessels are abnormal and fragile. They grow along the retina and along the surface of the clear, vitreous gel that fills the inner part of the eye. By themselves, these vessels do not cause symptoms or vision loss, but they have thin, fragile walls. If they leak blood, severe vision loss and even blindness can occur.

1.1 DR-related Lesions

Diabetic retinopathy is characterized by the presence of red (microaneurysms and hemorrhages) and bright (hard exudates, cotton wool spots) lesions as well as neovascularization. Drusen are also often observed in the retina, although they are associated especially with age-related macular degeneration (AMD) and can have similar appearance with the bright lesions [84].

This section explains the most common anomalies that are related to DR and can appear in eye-fundus images. Before enumerating the lesions associated with DR, please refer to Fig. 1.2 for an image depicting the main elements of the retina.

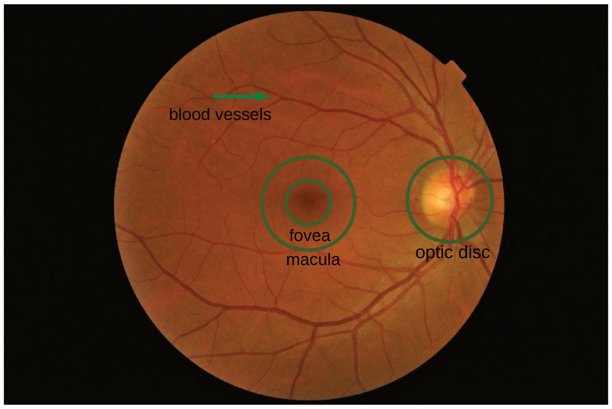


Figure 1.2: Example of a healthy retina with its typical anatomical elements.

- *Microaneurysms:* Consist of small outpouchings in capillary vessels which appear as small dots between the visible retinal vasculature [56]. Often occurs as one of the first signs of diabetic retinopathy [19].

- *Hard Exudates:* They are caused by the breakdown of the blood-retinal barrier, which leads to fluid rich in lipids and proteins to leave the parenchyma, causing retinal edema and exudation [72]. They have a yellow appearance and occur only in the occasional retinal image [19]. Fig. 1.3 exhibits an image with hard exudates.

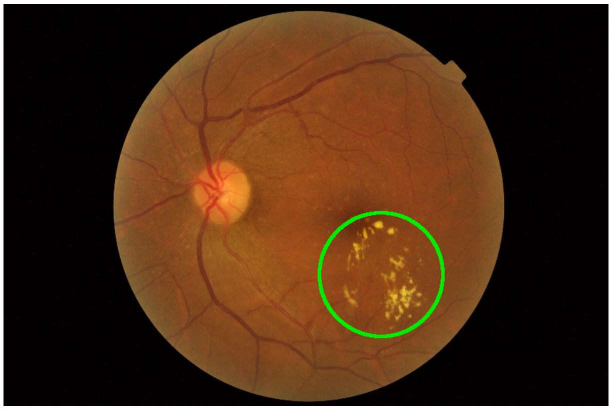


Figure 1.3: Example of a retinal image with Hard Exudates.

- *Hemorrhages:* They are similar to microaneurysms, but slightly larger and are found where capillary walls weaken. These may rupture causing intraretinal hemorrhages [72]. Superficial and deep hemorrhages are characterized as DR-related lesion. Fig. 1.4 shows an example of hemorrhages.

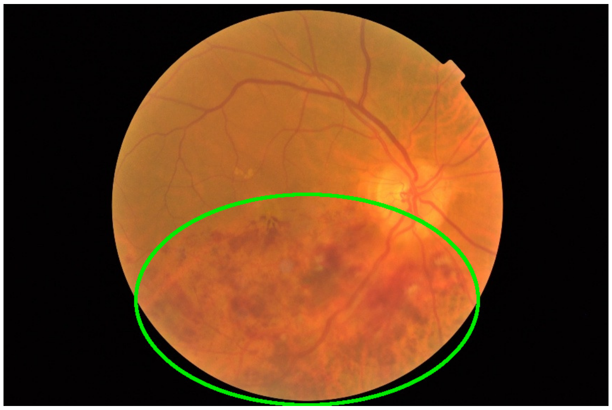


Figure 1.4: Example of a retinal image with Hemorrhages.

- *Cotton Wool Spots:* They appear as fluffy white patches on the retina and are caused by damage to nerve fibers. An image with cotton wool spots can be seen in Fig. 1.5.

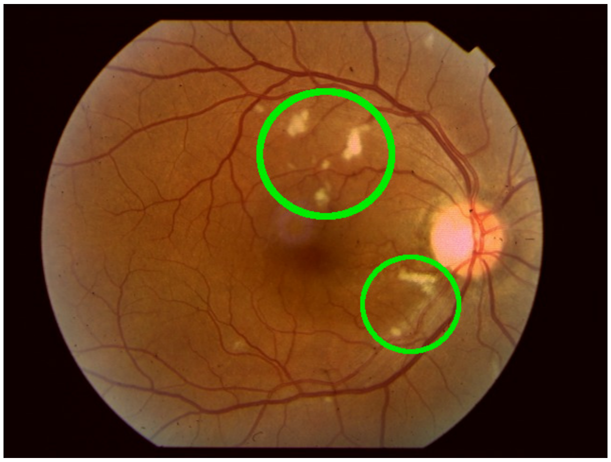


Figure 1.5: Example of a retinal image with Cotton Wool Spots.

- *Drusen:* They are bright lesions associated especially with age-related macular degeneration, which can have similar appearance, as well as from posterior hyaloid reflexes and flash artifacts, which can sometimes mimic bright lesions in appearance [57]. Fig. 1.6 depicts an image with hard drusen.

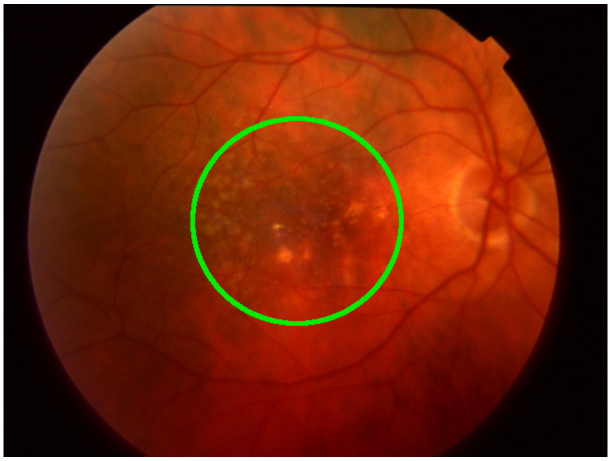


Figure 1.6: Example of a retinal image with Drusen.

- *Neovascularization:* The neovascularization process begins when it is detected the presence of intraretinal microvascular abnormalities. However, the new vessels are fragile and grow uncontrollably on the inner surface of the retina. An example of neovascularization is seen in Fig. 1.7.

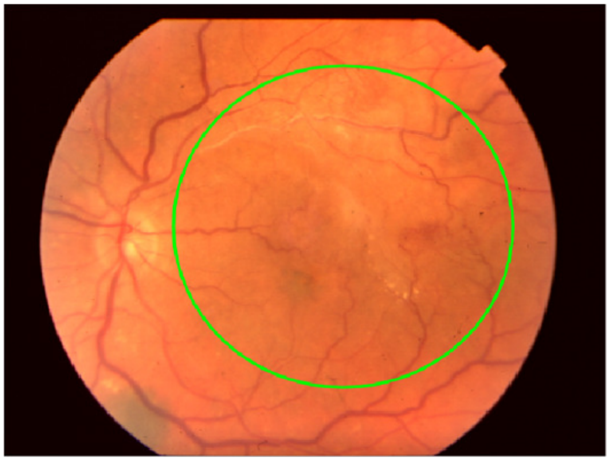


Figure 1.7: Example of a retinal image with Neovascularization.

1.2 Stages of the Work

This section introduces the four stages developed in this work and provides a sneak peek at the papers published and submitted for each one.

1.2.1 Quality Assessment

Image quality is an important aspect of automated image analysis and the factor that successful image analysis relies on. Although it is a common task in lesion detection projects, the manual quality assessment is expensive. Several works have discussed the assessment of image quality in the literature [20, 33, 45, 54]. However, most of the authors focus only on the *blur detection* (evaluating the presence of blurrings caused by motion) and discard important factors such as *field definition*.

For this stage, it was developed a method for analyzing image quality regarding motion blur and field definition. The work resulted in a paper entitled **Retinal Image Quality Analysis for Automatic Diabetic Retinopathy Detection** [65], published at the *XXV Brazilian Symposium on Computer Graphics and Image Processing* (*SIBGRAPI*), in 2012. The methods and results will be explained in Chapter 2.

Furthermore, it was also developed alternative methods for blur detection which will be described in Chapter 3. The resulting paper, entitled **Quality Control and Multi-lesion Detection in Automated Retinopathy Classification using a Visual Words Dictionary** [41], was accepted for publishing in the *35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (*EMBC*), in 2013.

1.2.2 Quality Improvement

Most of the works related to the DR detection apply diversified pre- or post-processing image techniques which ensure a correction in blurring artifacts. Furthermore, the use of image processing techniques improves considerably the results obtained in the classification step.

However, we discard the quality improvement task based upon the satisfactory results achieved for the quality assessment (see Chapters 2 and 3) and the facility in capturing new eye-fundus images and repeating the process in real-time.

1.2.3 DR-related Lesion Detectors

Due to several lesions related to DR and their diversified characteristics, there are several works present in the literature which focus on the detection of individual lesions, exploiting particular pre- and post-processing methods for each disease. In this stage, it was developed a series of individual detectors for the most important DR-related lesions: hard exudates, superficial hemorrhages, deep hemorrhages, cotton wool spots, and drusen. An additional classifier able to detect both superficial and deep hemorrhages was also implemented: red lesions.

Chapter 4 comprises the description of the experiments performed for the detection of individual DR-related lesions, as well as presents the experimental results for each anomaly. The development of DR-related lesion detectors represents an essential part of the paper entitled **Advancing Bag-of-Visual-Words Representations for Lesion Classification in Retinal Images**, submitted to a top-tier journal.

1.2.4 Detector Fusion

Given a set of detectors of individual DR-related lesions developed with a method which provides satisfactory results for the definition of presence/absence of the most common anomalies, this work involves the use of combining approaches aimed at pointing out whether an image is normal or has any lesion including possible ones not present during training.

The classifier fusion was explored for combination of the individual DR-related lesions and a paper, entitled **Data Fusion for Multi-lesion Diabetic Retinopathy Detection** [40], was published in the *25th IEEE International Symposium on Computer-Based Medical Systems* (*CBMS*), 2012. Chapter 5 contains the explanation of the methods used for fusion and results of the cited paper, as well as details and results of more recent experiments of this work.

1.2.5 Referral

In order to achieve early detection of DR, helping to stop or slow down its progress, international guidelines recommend annual eye screening for all diabetic patients. However, the increasing number of diabetic patients and the decreasing number of ophthalmologists make this suggested annual examination difficult to be performed sufficiently [22]. This factors tend to overwhelm the specialist even more during the next years.

Thus, aiming at referring to a specialist only the patients who really need a consultation, this work includes a stage for classifying retinal images as referable (to be referred to a specialist) or non-referable (not to be referred to a specialist) in the interval of one year. The methods, experiments and results are presented in Chapter 6 and were submitted as a paper entitled **Assessing the Need for Referral on Diabetic Retinopathy Treatment** to a top-tier journal.

1.3 Overview

Fig. 1.8 depicts an overview of this work. The first step consists of the quality analysis (Chapters 2 and 3). The second step focuses on the detection of individual DR-related lesions (Chapter 4). The third step receives the classification scores extracted in the second stage to create a multi-lesion detection framework (Chapter 5). Finally, similar to the third step, the fourth one explores the classification scores to assess the necessity of referral for a patient (Chapter 6).

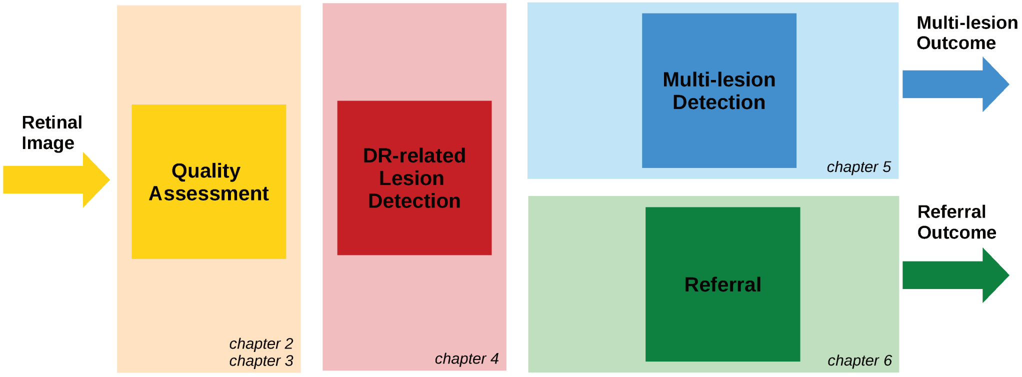


Figure 1.8: Overview of the work.

Chapter 2

Retinal Image Quality Analysis for Automatic Diabetic Retinopathy Detection

Given that sufficient image quality is a necessary prerequisite for reliable automatic detection systems in healthcare environments, the first step of this project is the assessment of retinal image quality. In this chapter, we present the methods employed for the quality assessment of retinal images and present the achieved results. The methods developed herein resulted in the publication [65].

2.1 Preamble

Diabetes and associated complications including diabetic retinopathy (DR) is increasing with a predicted prevalence tripling by 2050 in the United States [74]. Developing countries and Indigenous populations are likely to exceed this percentage [81]. In addition, DR is the leading cause of blindness in developed countries and therefore screening and targeted case management programs that are economically viable and identify and implement early treatment are required [61].

Mobile screening of high-risk populations, especially in rural and remote locations is an effective means of increasing the screening coverage of DR prevention programs [14]. Two-field photography in the hands of photographers with diverse skill levels and irrespective of using mydriatic or nonmydriatic photography compares favorably to ophthalmic investigations by specialists in metropolitan clinics [50].

To further enhance rural and remote area screening, automated image analysis programs have been developed and are now in use as a first line screening for microaneurysms in Scotland [63]. Several algorithms have been proposed for detecting parts of the retina, the presence/absence of retinopathy as well as specific lesions from mild nonproliferative to proliferative retinopathy and maculopathy (see [38] and references therein). An important aspect of automated image analysis and the factor that successful image analysis relies on is *image quality*.

Assessing image quality has been discussed in the literature by a number of authors [20, 33, 45, 54] and represents an important limiting factor for automated DR screening [59]. Image quality is reduced by artifacts in the image such as eye lashes or dust specs on the lens, only part of the retina is seen, the image is out-of-focus or the image is badly illuminated or blurred, among others. Image compression is often included with current software packages, which affects quality as does the resolution, field of view and type of camera [20]. Not directly related to image quality is retinal epithelial background, which often makes microaneurysm detection more difficult if the classifier is not trained for the specific ethnic group [42].

Furthermore, to ensure that automatic screening will be able to identify lesions like deep and superficial hemorrhages, it is necessary that the retinal images cover the appropriate portion of the retina, making the blood vessels visible. According to [26], the photographs should be centered on the macular region (See Fig. 1.2). Some authors have analyzed this aspect of image quality, known as *field definition* [28].

This chapter proposes methods for verifying these important factors of retinal image quality: *field definition* and *blur detection*. We aim at finding approaches that work well especially when trained with one type and tested with other types of retinal images. By introducing and adapting techniques such as visual words, quality analysis by similarity measures and classifier fusion to this context, we achieve promising classification results. In particular, for the field definition, our method is able to accurately distinguish between appropriate and inappropriate retinal images for automated DR screening.

2.2 Related work

Several methods for retinal image quality analysis are based on edge intensity histograms or luminosity to characterize the sharpness of the image [45]. In both approaches, the quality of a given image is determined through the difference between its histogram and the mean histogram of a small set of good-quality images used as reference.

Retinal morphology-based methods such as detection of blurring and its correlation to vessel visibility and retinal field definition have been applied for automatic detection of retinal image quality [33, 29]. The method of image assessment proposed by Fleming et al. [29], similarly to our work, involves two aspects: (1) image clarity and (2) field definition. The clarity analysis is based upon the vasculature of a circular area around the macula. The authors concluded whether or not a given image has enough quality using the presence/absence of small vessels in the selected circular area as evidence. The approach proposed by Fleming et al. requires a segmentation step to find the region of interest. However, for low-quality images, detecting segmentation failures is trivial.

Niemeijer et al. [54] proposed a method for image quality verification that is comparable to the well-known visual words dictionary classification technique, used extensively in pattern recognition tasks [89] and also one of the methods we rely upon in this work and in the following chapters. The purpose of Niemeijer et al. was to identify image structures that were present in a set of images. Local image structure at each pixel is described using the outputs of a set of 25 filters. Because the raw features are too numerous to be used directly in the classification process, a clustering algorithm is used to express the features in a compact way creating a visual dictionary. Once the visual dictionary is built, the features of each pixel are mapped onto words and a histogram of word frequencies for each image is created. These histograms are used to feed a classifier.

Visual words dictionaries constitute one of the approaches proposed to analyze image quality in this work. However, different to [54] we utilize visual words in the space of features representing discontinuities in the retina and not directly on every pixel. Second, our method is based on points of interest which are reasonably robust to some image distortions (e.g., rotation) and exhibit high repeatability, which allows us to easily find similar discontinuities in different images. Third, we have used the same method to detect lesions associated with DR in another work of ours [40]. Finally, the visual words dictionary calculated on the space of features exploits the benefits of an all-in-one classification algorithm which does not require any pre- or post-processing of the image.

Although good results for the assessment of diabetic retinal image quality have been obtained previously, the authors have not paid attention to one crucial factor needed for an acceptable screening of diabetic retinopathy. The image has to encompass the correct portion of the retina [26]. An analysis of DR images can fail because of inadequate field definition. As one exception, Fleming et al. [29] reported retinal image field definition in their work. In the viewpoint of the authors, an image is defined as having adequate field definition if it satisfies a series of constraints, that aim at verifying distances between important elements of the anatomy of the retina, such as the optic disc and fovea (Fig. 1.2).

2.3 Technique for Field Definition

Here, we discuss a simple method for verifying the field definition. In this problem, a good retinal image for further DR analysis is one image centered on the macula (See Fig. 1.2).

The method we discuss here operates based on the methodology of full-reference comparison. In this methodology, a reference image with assured quality is assumed to be known and quantitative measures of quality for any image are extracted by comparisons with the reference [86]. Given that the macular region has a distinguishable contrast in comparison with the remaining regions, and we are interested in the content of the center of retinal images, metrics of similarity have shown to be highly suitable for this objective.

We selected a set of images centered on the macular region as well as a set of images not centered on the macular region (centered on the optic disc or in any other location on the retina). Then, we calculated similarities between a given image and the reference images (positive and negative), with respect to their central regions and created a feature vector for later classification. In the next section, we explain the method employed for the feature extraction as well as the learning step of the technique for field definition.

2.3.1 Characterization

Wang et al. [86] proposed a new philosophy for comparison of images that considers image degradation as perceived changes in structural information instead of perceived errors (visibility of errors). The method, known as Structural Similarity (SSIM) [86] is calculated according to Eq. 2.3 which we shall define later.

Given that we are interested in assessing if the macula is present in the center of the image and it is clearly different from other regions of the retina, we use one region of interest (RoI) of pre-defined size (121 × 121) on the center of the retinal image. Fig. 2.1 depicts some positive (centered on the macular region) and negative (centered on the optic disc or in other region) RoIs.

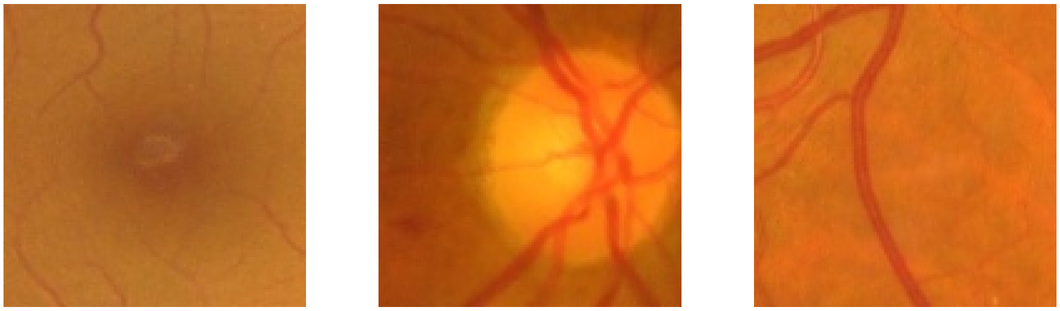


Figure 2.1: Examples of RoIs whose images are centered on the macula (left), centered on the optic disc (middle), and non-representative (right).

To characterize each retinal image, we measure the structural similarity between the RoI of the image of interest and the RoIs of a set of reference images and calculate their average. We selected a set of 40 retinal images for reference (20 represent the retina with good field definition and 20 that would be discarded for not being centered on the macula). For the group not centered on the macula, we selected 12 RoIs centered on the optic disc and eight in any other area. The reference images are not used further neither for training nor for testing.

As we are comparing pixels directly, we investigated if a simple contrast normalization technique helps to boost classification results. For that, we tested the use of the images in grayscale as well as in RGB color space with and without the normalization considering contrast limited adaptive histogram equalization (CLAHE) [66]. CLAHE is suitable to improve the local contrast of an image.

After comparing each image with the references, its feature vector considering color images comprises 18 features: three comparison functions from SSIM × three color channels (RGB) × two sets of reference patches (positive and negative). SSIM was calculated breaking Eq. 2.3 to three terms: luminance, contrast, and structure according to [86].

2.3.2 Learning

At the end of the characterization process, we have a set of feature vectors representing the structural similarities with positive and negative reference images. The final classification procedure is performed using the *Support Vector Machine* (SVM) algorithm [16]. Although we could use other classifiers, we opted for SVM classifiers for a number of desirable traits: their solutions are global and unique; they have a simple geometric interpretation; and they do not depend on the dimensionality of the input space.

We train the classifier with feature vectors calculated from training images containing positive (images centered on the macular region) and negative (images centered on any other region of the retina) examples. When training the SVM, we use "grid search" for fine tuning the SVM parameters based only on the training examples [16].

2.4 Technique for Blur Detection

Although image quality analysis can have several ramifications before arbitrating on the quality of an image, we focus on two very common problems during image acquisition: blurring and out-of-focus capture.

2.4.1 Characterization

The method involves a series of different blurring classifiers and classifier fusion to optimize the classification. Next, we present the details of the methods we use for blurring classification. Basically, we rely upon four descriptors: vessel area, visual dictionaries, progressive blurring and progressive sharpening. We also explore combinations of them.

Area Descriptor Given that blurring affects the visibility of the blood vessels, our first descriptor consists of the measurement of the area occupied by the retinal vessels. For that, we calculate the image's edge map using the Canny algorithm [36]. Next, we measure the area occupied by the vessels counting the quantity of pixels on the edges and dividing it by the retina's total number of pixels. Fig. 2.2 depicts retinal images followed by their respective Canny edge maps.

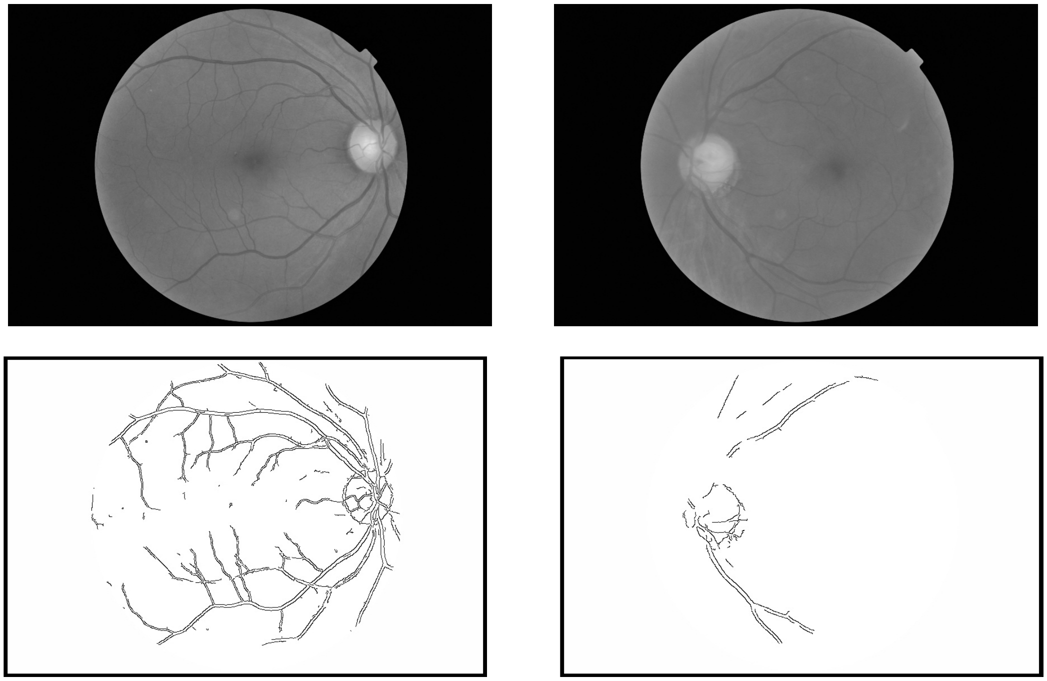


Figure 2.2: Retina with enough quality (left) and with blurring (right) with their respective Canny edge maps (inverted for visualization purposes).

In the end of the characterization phase, we have an 1-d feature vector whose area descriptor is the unique feature.

Visual Dictionary Descriptor In this descriptor, each image is characterized by finding stable points of interest (PoIs) across multiple image scales that capture image discontinuities. We are interested in characterizing an image in order to capture any inconsistencies/discontinuities it might have (e.g., blood vessels) in order to classify it.

To build a visual dictionary and define whether a specific retinal image has enough quality, training images tagged as having quality (no blur) by a medical specialist as well as images associated with blurring are required. After collecting the training images, the next step consists of finding the points of interest in all training images. To detect the points, we use the *Speeded Up Robust Features* (SURF) [10] as it is a good feature detector with reasonable speed.

From the points of interest representing the images with quality as well as blurred images during training, we randomly select a set of PoIs for each group. At this stage, the number of PoIs (k) to be retained as representative of the quality or non-quality images is decided. We find the $k/2$ points of interest associated with a high-quality image and repeat the process to find the $k/2$ points associated with images with blurring. We refer to these k points of interest as a visual dictionary. Note that this is different from other approaches in the literature (e.g., [78, 25]) which normally find a global unique dictionary and not one per class. In our experience, class-aware dictionaries are more appropriate for retinal images. The *class-aware* treatment is explained in detail in Chapter 4

In order to use any machine learning method, the next step is to map the PoIs within each image to the most representative points in the dictionary. For each image, we associate each one of its PoIs to the closest word in the dictionary using Euclidean distance. In the end, each training image is represented by a histogram of k bins which counts the number of times each PoI in the image was mapped to that word in the dictionary. We used such histogram as the image's feature vector. During testing, the process is simple: we extract the points of interest of the test image and map its PoIs to the dictionary creating its k dimensional feature vector.

Determining the optimal number of clusters for any given set is still an open problem and is therefore best determined empirically. In our experiments, we evaluated the performance of the visual dictionary descriptor with k = 30, 50, 70, 100 and 150. We avoided bigger dictionaries in order to keep the classification process fast and accurate. The visual dictionary approach is described in detail in Chapter 4.

Blurring, Sharpening, Blurring + Sharpening descriptors We propose a variation of the traditional method widely employed in the literature to quantify the visibility of errors: full-reference method for assessment of quality [86]. In our variation, the reference image is not defined previously, but each image under analysis is elected as a reference and compared to progressive transformations of itself.

For the blurring descriptor, we progressively blur the input image with different intensities and measure how much the image can lose the discontinuities that characterize the blood vessels. It is expected that an image with poor quality be more similar to its transformed version than a good-quality image in comparison with its transformed version.

For the sharpening descriptor, we employ different sharpening filters that enhance edges and provide higher similarity values for good-quality images than for blurred images.

The sharpening filter is a simple sharpening operator which enhances edges (and other high frequency components in an image) via a procedure which subtracts a smoothed version of an image from the input image.

To explore simultaneously the two features, we investigated a Blurring + Sharpening descriptor to represent retinal images.

Each input retinal image is considered as a reference image and is compared with its filtered images. For that, we define a filter-bank as a set of rotationally symmetric Gaussian lowpass filters $G_\sigma(i,j)$. The set comprises 12 filters with kernel sizes $k_s \times k_s$ where $k_s \in \{3,5,7\}$, and standard deviations $\sigma \in \{0.5, 1.5, 3.0, 4.5\}$.

For the blurring descriptor, each resulting image $f^i_{smooth}(x,y)$ is a filtered version of the original image $f(x,y)$, denoted as

$$f^i_{smooth}(x,y) = \sum_{i,j}^{k_s} G_\sigma(i,j) f(x+i, y+j) \tag{2.1}$$

For the sharpening descriptor, each resulting image $f^i_{sharp}(x,y)$ is calculated as

$$f^i_{sharp}(x,y) = f(x,y) + \lambda(f(x,y) - f^i_{smooth}(x,y)) \tag{2.2}$$

where λ is a scaling constant $\in [0.0, 1.0]$. Here, we fixed the constant, $\lambda = 0.7$ without any further analysis.

For each retinal image, we measured the similarity between the input image $f(x,y)$ (considered as reference) and each response image $f^i(x,y)$ blurred or sharpened according to the descriptor of interest. We calculated the similarity $sim(f(x,y), f^i(x,y))$ using three different metrics:

- SSIM: the structural similarity index between two images can be viewed as a quality measure of one of the images being compared, provided the other image is regarded as of good quality. We calculated SSIM for 11×11 windows centered on every pixel. The result is a matrix with the same dimensions as the compared images. We report the final similarity value as the average of such matrix. The $SSIM(R,S)$ where R and S are two 11×11 windows centered on a pixel (x,y) is given by

$$\begin{aligned} SSIM(R,S) = {} & (2\mu_R\mu_S + c_1)(2\sigma_{RS} + c_2) \times \\ & 1/[(\mu_R^2 + \mu_S^2 + c_1)(\sigma_R^2 + \sigma_S^2 + c_2)] \end{aligned} \tag{2.3}$$

  where μ_R and μ_S are the average of R and S regions, σ_R^2 and σ_S^2 their variances, σ_{RS} their covariance, c_1 and c_2 are two variables to stabilize the division with weak denominator. These variables depend upon two constants $k \ll 1$ ($k_1 = 0.01$ and $k_2 = 0.03$) and the image's dynamic range L which is 255 in our case. The final values, for $c = (k * L)^2$, are: $c_1 = 6.5$ and $c_2 = 58.5$.

- SSD: the sum of squared differences is calculated by subtracting pixels between the reference image $f(x,y)$ and the target image $f^i(x,y)$. The differences are squared.

$$SSD(f(x,y), f^i(x,y)) = \frac{1}{MN} \sum_{x,y} [f(x,y) - f^i(x,y)]^2, \tag{2.4}$$

  where M and N are the number of rows and columns.

- NCC: the normalized cross correlation is defined as

$$NCC(f(x,y), f^i(x,y)) = \frac{1}{MN} \sum_{x,y} \frac{f(x,y) f^i(x,y)}{\sqrt{f(x,y)^2} \sqrt{f^i(x,y)^2}}, \tag{2.5}$$

  where M and N are the number of rows and columns.

For each image, the blurring and the sharpening descriptors have feature vectors with 108 similarity measures: 12 gaussian filters × 3 metrics of similarity × 3 color channels (RGB). The blurring + sharpening descriptor is the concatenation of the feature vectors extracted by the blurring and the sharpening descriptors leading to a 216-d feature vector.

2.4.2 Learning

In the end, for each retinal image, we have a set of five feature vectors considering the area descriptor, visual dictionary descriptor, blurring and sharpening descriptors and their concatenation. The final classification procedure is performed using the SVM algorithm [16].

We train the classifier with feature vectors calculated from training images containing positive (images tagged by a medical specialist as good quality) and negative (images tagged by a medical specialist as containing blur) examples. When training the SVM, we use "grid search" for fine tuning the SVM parameters based only on the training examples [16].

2.4.3 Fusion

It is possible that a series of complementary classifiers are more suited to accurately assess the quality of retinal images operating over several instances observed in the two classes of images. For example, analyzing not only one characteristic, but a series as the area occupied by visible blood vessels, the distributions of positive/negative visual words, similarities with blurred images and similarities with sharpened images provides a higher probability of correctly evaluating any retinal image from any camera.

We evaluated two approaches for fusion: at feature-level combining the feature vectors directly by concatenation and at classifier level by creating a Meta-SVM classifier (or meta-classification) trained over the outputs of individual classifiers, in this case, the marginal distances to the decision hyperplane produced by the SVMs.

2.5 Experiments and validation

This section shows the results for evaluating the quality of an image with respect to field definition and blurring artifacts as an effective pre-processing before using any classifier for detecting diabetic retinopathy lesions.

There are many metrics to measure the success of a detection/classification algorithm. For the purposes of this project, we are interested in *per image* metrics, such as sensitivity (number of images tagged as having enough quality over the total number of images with quality), and specificity (number of images tagged as blurred over the total number of blurred images). However, for quantifying the performance of the proposed methods, we calculated the area under the receiver operating characteristic curve (ROC). The area under the curve (AUC) is an accuracy measurement that explores how well the classifier is based on its ROC curve. An AUC of 100% represents a perfect test while an area of 50% represents a worthless test.

We organized the experiments in four rounds:

- **Round #1 – Single results for field definition.** Field definition approach using single classifiers. We performed all tests on single datasets using 5-fold cross-validation.

- **Round #2 – Cross-dataset results for field definition.** Cross-dataset approach, in which we trained the field definition classifiers in one dataset and test in another. We evaluated the ability of the field definition system to operate over images from different acquisition conditions.

- **Round #3 – Single results for blur detection.** Blur classification using single classifiers. We also evaluated fusion methods to check if they improved the classification results. We performed all tests on single datasets using 5-fold cross-validation.

- **Round #4 – Cross-dataset results for blur detection.** Cross-dataset approach, in which we trained the blur classifiers in one dataset and tested in another. We evaluated the ability of the blur classifiers to operate over images from different acquisition conditions.

In the 5-fold cross-validation protocol, we split the dataset into five parts, train with four parts and test on the fifth, repeating the process five times each time changing the training and testing sets.

2.5.1 Datasets

We performed the experiments for quality analysis using the DR1 and DR2 datasets annotated by medical specialists.

The DR1 dataset is from the ophthalmology department of Federal University of São Paulo (Unifesp), collected during 2010. It comprises 5,776 images with an average resolution of 640 × 480 pixels. 1,300 images have good quality (do not contain blur and are correctly centered on the macula), 1,392 represent poor quality (blur) and 3,084 are diagnosed as images of the periphery (not centered on the macula). Three medical specialists manually annotated all of the images. The images were captured using a TRC-50X (Topcon Inc., Tokyo, Japan) mydriatic camera with maximum resolution of one megapixel and a field of view of 45 degrees.

The DR2 dataset is from the same ophthalmology department, collected during 2011. One medical specialist graded the images. DR2 comprises 920 12.2MP images decimated to 867 × 575 for speed purposes and containing 260 images not centered on the macula (146 centered on the optic disc and 114 not centered on any interesting region) and 660 images centered on the macula (466 good and 194 low quality). The images were captured using a TRC-NW8 retinographer with a Nikon D90 camera.

For more details and for downloading the datasets, please refer to `http://www.recod.ic.unicamp.br/site/asdr`.

2.5.2 Round #1: Single Results for Field Definition

Here, we explore the measures of structural similarity in order to create a classifier able to analyze a retinal image and evaluate if it comprises the correct portion for diabetic retinopathy screening (centered on the macula).

We performed four experiments for field definition. In the first experiment, the images were analyzed in grayscale. The second experiment also was performed with the images in grayscale, but after an adaptive histogram equalization (CLAHE). Next, we considered the case of color images with and without histogram equalization.

For all experiments of field definition, we used 40 reference images. All of them were not considered further for training nor for testing.

Fig. 2.3 and Fig. 2.4 depict the ROC curves for the field definition approach using 5-fold cross-validation protocol of the DR1 and DR2 datasets, respectively.

As we can observe in Fig. 2.4, the method achieves reasonably successful results for field definition. The experiments using the DR2 dataset present even better results. The experiment with color images considering histogram equalization provides the best result, but this result in not statistically different to the others in DR2. However, in the experiments using the DR1 dataset (Fig. 2.3), that comprises a larger quantity of images (1,300 positives and 3,084 negatives), we can note a great difference of AUCs between the different techniques. The method that uses the color images without requiring an adaptive histogram equalization is the highlight.

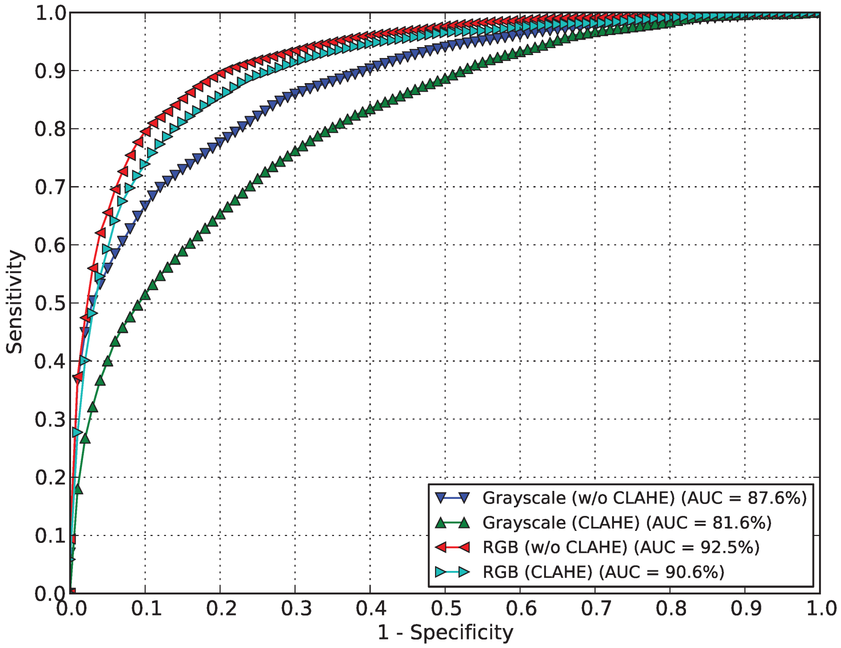


Figure 2.3: DR1 field definition using 5-fold cross-validation.

As mentioned, there is not a considerable difference between the experiments with and without adaptive histogram equalization using the DR2 dataset. The reason is that the images from DR2 present small variations in illumination. The images from DR1 dataset exhibit a high variation of illumination making the CLAHE insufficient to distinguish them and improve classification.

2.5.3 Round #2: Cross-dataset Results for Field Definition

Conventional detectors usually build a classifier from labeled examples and assume the testing samples are generated from the same distribution. When a new dataset has a different distribution from the training dataset (e.g., different acquisition conditions), the performance may not be as expected.

In this round, we validated the field definition approaches considering the problem of cross-dataset field definition testing, which aims at generalizing field definition models built from a source dataset to a target dataset. We refer the DR1 as the source dataset (training), and the DR2 as the target dataset (testing). We emphasize that the two datasets were collected in very different environments with different cameras, at least one year apart and in different hospitals.

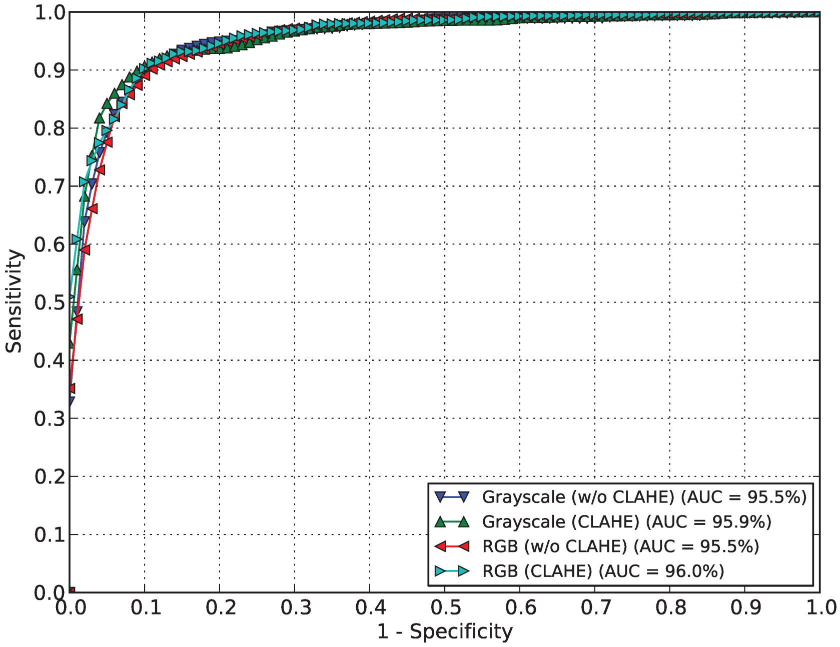


Figure 2.4: DR2 field definition using 5-fold cross-validation.

Table 2.1: Field definition: AUC for the experiments.

Method	DR1	DR2	Cross
Grayscale	87.6%±0.7%	95.5%±1.3%	84.7%
Grayscale (CLAHE)	81.6%±0.6%	95.9%±1.2%	83.2%
RGB	92.5%±0.7%	95.5%±1.1%	75.5%
RGB (CLAHE)	90.6%±0.9%	96.0%±0.8%	75.6%

For this round, we trained the classifiers with DR1 dataset (3,064 images located on the periphery of the retina, 1,280 images centered on the macula and 40 images removed and used as reference), and tested with DR2 dataset (260 images not centered on the interest region and 660 images centered on the macular region).

Fig. 2.5 presents the ROC curves achieved by the method under the cross-dataset validation.

As discussed in the previous section, the high variation of the illumination in DR1 in comparison with DR2 makes the histogram equalization technique unable to improve the results. Table 2.1 summarizes the results for field definition for the single and cross-dataset tests.

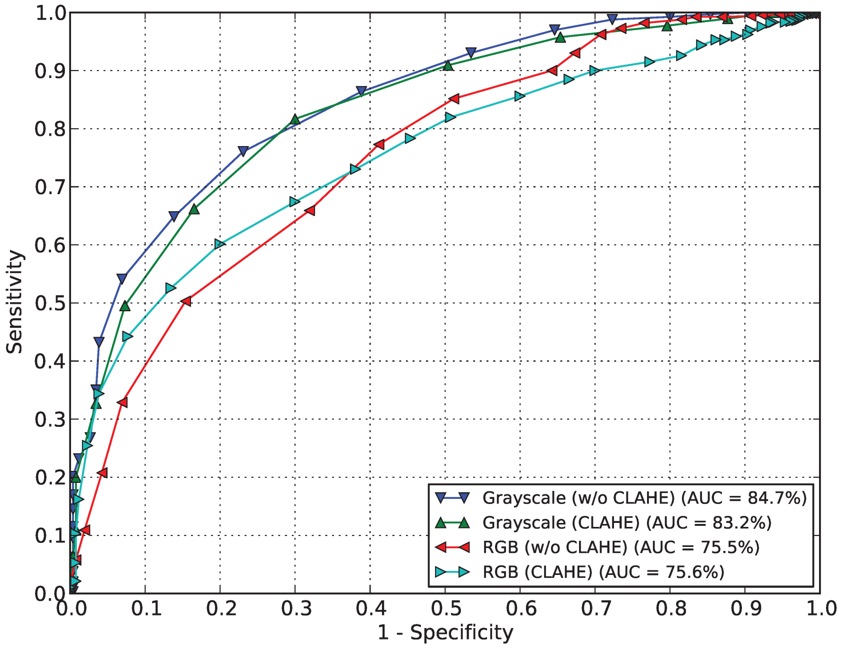


Figure 2.5: Cross-dataset validation for field definition using DR1 as training and DR2 as testing sets.

Comparison with State of the Art

In a previous work, Fleming et al. [29] introduced the first automatic field definition study. The authors obtained 95.3% for sensitivity and 96.4% for specificity. Our results for field definition are somewhat comparable to the previous results (96% AUC, and 93% sensitivity and 92% specificity using DR2 and RGB-CLAHE). However, Fleming at al. used a different dataset with 1,039 retinal images and did not evaluate the algorithms in a cross-dataset scenario.

2.5.4 Round #3: Single Results for Blur Detection

In the third round, we performed experiments to verify the descriptors and classifiers to separate good-quality images from blurred ones. We explored several descriptors, each one trying to take full advantage of the differences observed between poor and good-quality images, aimed at providing a series of blur classifiers. In this experiment, we developed classifiers that work in parallel, assuming competitive operation and contributing equally to the final decision.

Fig. 2.6 and Fig. 2.7 depict the results for DR1 and DR2 datasets.

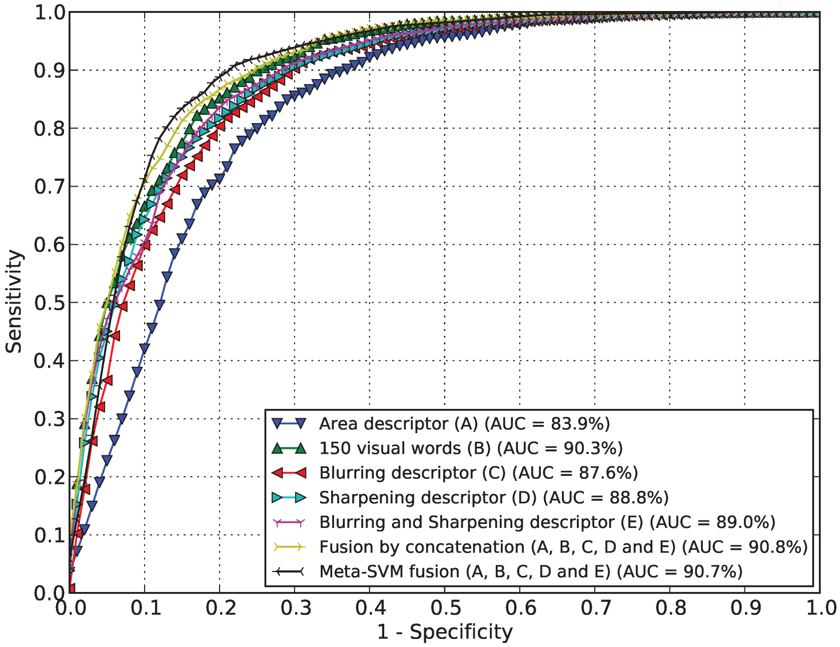


Figure 2.6: DR1 blur classification using 5-fold cross-validation.

Table 2.2 summarizes the results. The ROC curves as well as the areas under the curves reflect that interesting results are obtained for blur classification. We observe in the table that, for single classifiers, the best result using the DR1 dataset was achieved by the visual words approach (a dictionary size of 150 words was previously defined as the best number of words for the dictionary and not shown here). For the DR2 dataset, the visual words approach also presents good results but are outperformed by classifiers trained with the blurring and sharpening descriptors. The blurring, sharpening and the blurring + sharpening descriptors provide acceptable results in both datasets.

As expected, the more exciting results were provided by the fusion methods. As discussed before, exploring not only one evidence of incoherence, but several complementary information of poor and good-quality images, gives more chances of obtaining better results. In our case, the ensemble method that uses only the concatenation of the feature vectors provides the highest result for DR1 (AUC = 90.8%), followed closely by the Meta-SVM fusion method (AUC = 90.7%).

Here, it is important to emphasize that the ensemble by concatenation operates on large feature vectors making the method highly sensitive to the curse of dimensionality,

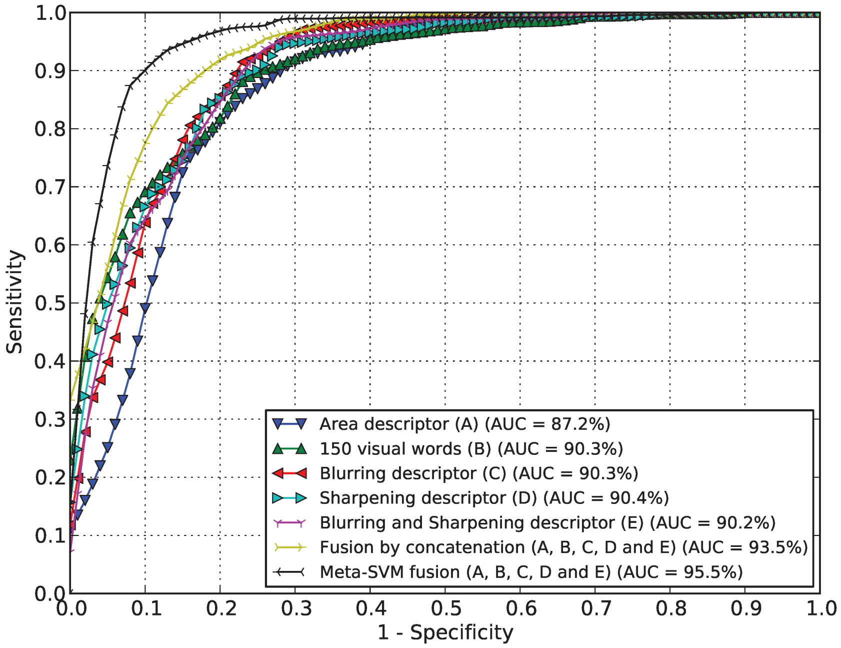


Figure 2.7: DR2 blur classification using 5-fold cross-validation

and presents limitations for classification for specific classifiers and specific machines [72]. In addition, it is often necessary to deal with complicated normalization techniques to put different features in the same domain [72]. Conversely, the Meta-SVM fusion method is less subject to such limitations, since it only adds a new level of classification on a response vector composed of five classification scores (distances to the decision hyperplane) provided by the individual classifiers.

For the DR2 dataset, the highest AUC was obtained with a large difference using the Meta-SVM fusion method (AUC = 95.5%), followed by the fusion by concatenation technique (AUC = 93.4%).

2.5.5 Round #4: Cross-dataset Results for Blur Detection

The last round of experiments explored the cross-dataset validation to evaluate how the classifier models built from a source dataset (DR1) to a target dataset (DR2) generalize.

For this round, we trained the classifiers with DR1 (1,392 images with poor quality and 1,300 images with good quality) and tested the classifiers with DR2 dataset (194 retinal images with enough quality and 660 images with no quality). Fig. 2.8 depicts the resulting ROC curves.

Table 2.2: Blur Detection: AUC for the experiments.

Descriptor/Fusion	**DR1**	**DR2**	**Cross**
Area	83.9%±2.4%	87.2%±2.6%	87.1%
Visual words	**90.3%±1.2%**	**90.3%±2.3%**	**85.6%**
Blurring	87.6%±1.3%	90.3%±2.6%	60.8%
Sharpening	88.8%±1.4%	90.4%±3.9%	83.9%
Blurring and Sharpening	89.0%±0.9%	90.2%±3.0%	69.0%
Fusion by Concatenation	**90.8%±0.9%**	**93.5%±1.4%**	**87.0%**
Fusion by Meta-SVM	**90.7%±2.3%**	**95.5%±1.6%**	**87.6%**

Observing the AUCs in Fig. 2.8 and summarized in Table 2.2, we note that the visual words descriptor presents satisfactory results using the cross-dataset protocol. However, the simple area descriptor is the highlight in this experiment, showing that the density of blood vessels may be considered as an acceptable approach to assess the quality of retinal images.

Fortunately, with this experiment we can show the importance of a cross-dataset validation protocol. Although the blurring descriptor showed interesting results in the validation with single datasets, here it failed along with blurring + sharpening combination. With them, a large number of images from the DR2 dataset was classified at the same distance to the SVM decision hyperplane. This fact happens because the DR1 has greater contrast and illumination variation than DR2 dataset and, therefore, the descriptions of the DR2 match to approximate scores given by a classifier trained with DR1. Consequently, a small amount of operating points are available, as we can see in Fig. 2.8. This effect might be reverted using image normalization techniques more complex than CLAHE but we did not investigate this in this chapter.

As we expected, detector fusion with the Meta-SVM method provides the best AUC with the caveat that in this analysis the Meta-SVM results are not statistically better than the single classifier using the single area descriptor.

Comparison with State of the Art

Our results are comparable to several prior results. The approach proposed by Niemeijer et. al. [54] and explained in Sec. 2.2 provided an AUC of 99.6% operating over a dataset comprising 1,000 images. Davis et. al. [20] achieved a sensitivity of 100.0% and a specificity of 96.0% using a dataset comprising 2,000 images. However, no conclusion can be drawed observing only the final results, since we must consider that the datasets are different (camera model, acquisition conditions) and the methodologies employed are distinct. We emphasize that only one dataset is not enough as a validation protocol for a reliable system.

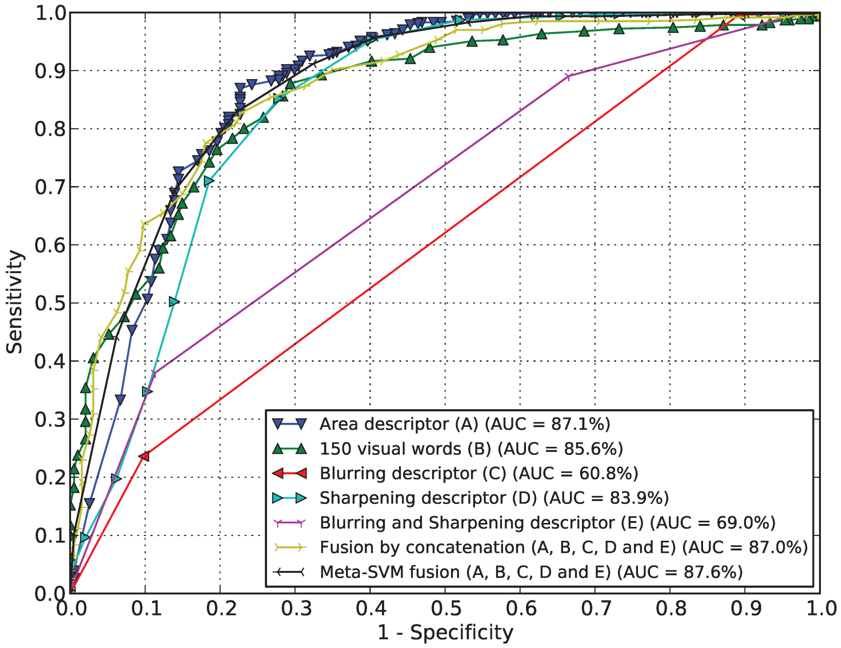


Figure 2.8: Cross-dataset validation for blur classification using DR1 as training and DR2 as testing sets.

2.6 Final Remarks

The assessment of diabetic retinal image quality presented in this chapter shows promising results. Several studies have obtained satisfactory results for image quality verification in the literature. However, these have only focused on image quality as a generalized approach and have not paid attention to field definition, which is one crucial factor for an effective automatic screening of diabetic retinopathy. In addition, cross-dataset validation is hardly performed.

In the approach we discuss in this chapter, image quality was defined by two aspects: field definition and blur analysis. For field definition, we proposed the use of structural similarity measures to evaluate the quality of retinal images. We obtained an AUC of 96.0% using color images and the DR2 dataset.

For blur analysis, we explored several descriptors, each one taking full advantage of the specific variations between poor and good-quality images. Furthermore, we aimed at providing a series of blur classifiers that work in parallel, assuming competitive operations and contributing equally to the final decision. We also evaluated the use of fusion techniques and the best result was reached with the Meta-SVM fusion method (AUC =

95.5% on DR2 dataset).

With the proposed methods for assessment of diabetic retinal images, it is possible to devise and deploy a system capable of robustly identifying images with low quality and, afterwards, discard them. A retinal camera equipped with quality assessment methods would be adequate to analyze eye-fundus images taken in real-time, preventing misdiagnosis and posterior retake.

Chapter 3

Quality Control and Multi-lesion Detection in Automated Retinopathy Classification

In this chapter, we present another approach employed in this work for quality assessment. The methods employed herein resulted in the publication [40]. Although the paper also involves the detection of DR-related lesions, this chapter is limited to the quality evaluation.

3.1 Preamble

Machine learning methods and automated data mining are important for health informatics and have been actively investigated in automated classification of disease, including diabetic retinopathy [32, 35, 17, 79, 69]. Quality control is an important part of automated image analysis [28, 65] as is the detection of multiple lesions in images of different resolutions and ethnic background. This requires algorithms that unify image quality assessment and do not require preprocessing for each type of lesion separately, have a high accuracy for each type of lesion and, if possible, improve the classification when lesion types are combined in the classification framework. In this context, we have previously shown that visual word dictionaries have good accuracy with training of the classifier on different images to the test images and no preprocessing of the test images used in the research [42]. This chapter describes further developments using visual word dictionaries by considering a means of identifying poor quality images.

The rest of the chapter is organized as follows. Section 3.2 presents our method of visual word dictionaries adapted to determine the quality of an input image. Section 3.3 presents the results for the proposed approach in terms of image quality analysis. Finally, Section 3.4 concludes the chapter

3.2 Proposed Methodology

The contribution of this chapter is the proposal of the adaptation of the visual words dictionary methodology to classify whether or not an input image meets the quality standard required for automatic assessment. Although image quality analysis can have innumerable ramifications before arbitrating on the quality of an image, in this chapter we focus on a very common problem during image acquisition: *blurring*.

3.2.1 Quality selection

Among all types of problems associated with the image acquisition process, one of particular interest is the detection of blurred images. This chapter focuses on classifying the quality of an image based on blurring.

For this intent, the general visual words methodology, which was explained in Chapter 2 and whose formal definition is given in Chapter 4, needs to be adapted in order to capture an important particularity for retinal images: high-frequency information is more pronounced in the border regions associated with the venous branching pattern.

To capture the behavior such as blurring, the edge map of each training image is first calculated using the Canny algorithm [36]. Next, the representative patches for the image are centered using the edge map. Fifty non-overlapping patches (each one with 50×50 pixels) in the edge map are centered in order to capture the differences of such regions. We analyzed several sizes and quantity of patches and noted that 50 patches of 50×50 pixels were satisfactory to cover the edges of the blood vessels. The use of patches is the notable difference with respect to the general methodology described in Chapter 2. SURF is therefore not used directly on the image, rather it is directed to regions on the edge map that are more important to differentiate blur and non-blur artifacts, namely regions with edges.

After calculating the points of interest within the selected 50 regions, the most representative PoIs have to be found for each training image. For that, K-Means clustering algorithm is applied to select a specialized visual dictionary for image quality analysis. In this case, it is selected $k/2$ regions that represent good quality images and $k/2$ regions for low quality images. Fig. 3.1 depicts an example of a retinal image and its Canny edge map with the 50 patches centered on the localized edges. After generating each image feature vector, it is normalized using the traditional term-frequency (divide the entries by total sum of the bins).

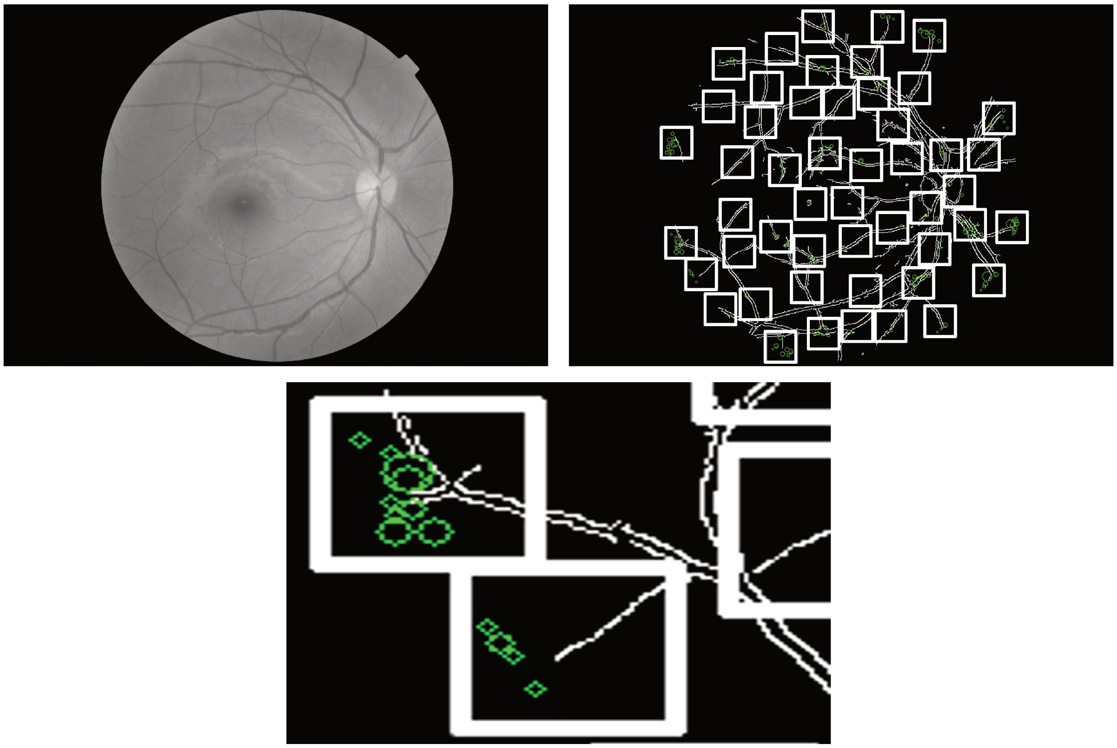


Figure 3.1: Input image with its Canny edge map as well as the 50 50×50-selected image regions centered on the edges (small squared regions) and the calculated SURF PoIs within each region (green circles), followed by a highlighted patch.

3.3 Results

This section shows the results for evaluating the quality of an image for automatic screening. All the experiments reported herein consider a 5-fold cross-validation protocol in which the data set is divided into five parts, train with four parts and test on the fifth, repeating the process five times each time changing the training and test sets.

3.3.1 Dataset

The experiments were conducted on the DR2 dataset from the Ophthalmology Department of the Federal University of São Paulo for which we have quality assessment grading performed by one medical specialist. DR2 comprises 660 12.2MP images decimated to 867 for speed purposes divided into 466 good and 194 low quality images captured using a TRC-NW8 mydriatic camera with a D90 camera for image capture. For more details and for downloading the data set, please refer to `http://www.recod.ic.unicamp.br/site/asdr`.

3.3.2 Image Quality

Fig. 3.2 depicts the results for image quality analysis. In this case, a good-quality image is one with no blurring. Note that the dictionary needs 70 words for a reasonable performance resulting in an AUC of 87.4%, in this case. For a dictionary with 30 words, the AUC is 86.4% while for 50 words the AUC is 85.7% and for 100 it is 81.1%. These are promising results, considering that this was a first attempt for solving image quality assessment, and that explores only one approach.

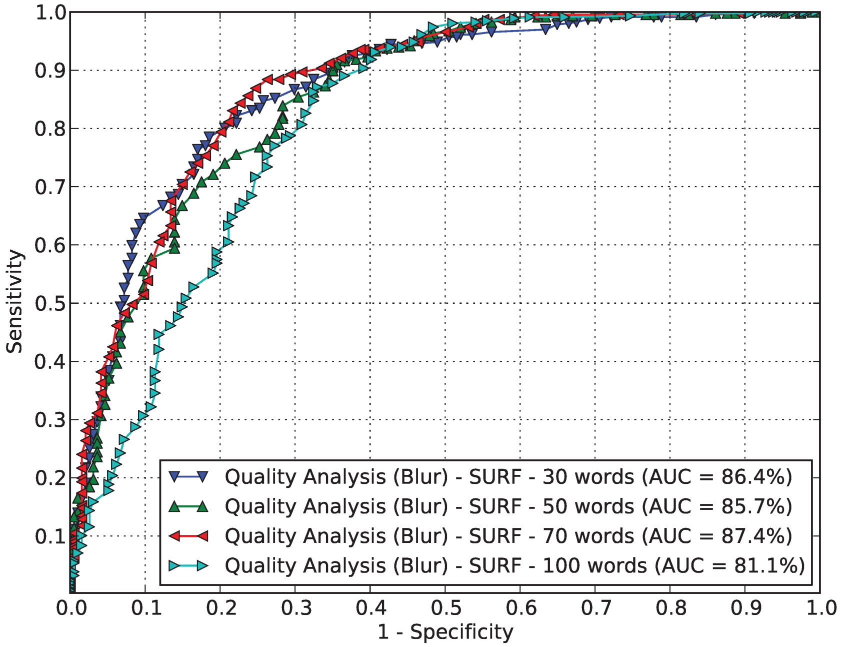


Figure 3.2: Image quality analysis considering 50, 50×50-regions per image from DR2 data set and various dictionary (most representative regions) sizes.

Table 3.1 shows, for comparison purposes, more recent experiments previously presented in Chapter 2. The column brought to this chapter presents the results achieved with the same dataset and validation protocol. We can observe that extraction of local features in patches on edge maps outperforms only the experiment with Area descriptor.

Table 3.1: Blur Detection: AUC for more recent experiments.

Descriptor/Fusion	**DR2**
Area	87.2%±2.6%
Visual words	**90.3%±2.3%**
Blurring	90.3%±2.6%
Sharpening	90.4%±3.9%
Blurring and Sharpening	90.2%±3.0%
Fusion by Concatenation	**93.5%±1.4%**
Fusion by Meta-SVM	**95.5%±1.6%**

3.4 Final Remarks

Many feature descriptors have been proposed in the literature for problems like copy detection [83] or object localization [77], for example: Gaussian derivatives [31], complex features [9], SIFT [48], and SURF [10]. Such methods need to capture sufficient image details, whilst being robust to small deformations or localization errors [10]. Using the Hessian approximation within the visual word dictionary framework is comparable to and, in some instances, better than current state-of-the-art interest point detectors. SURF's advantage relies on its robustness against rotation, scale change, image noise, change in brightness across the image and change of view being suitable for adaptation for a classification framework instead of its usual image matching form.

The extraction of local features in regions associated with the venous branching pattern yielded promising results for analyzing retinal image quality. However, more recent experiments, whose results are showed in Table 3.1 and methodology was explained in Chapter 2, showed that the use of edge maps was not suitable for this goal. The table presents results which outperform this one (except for Area descriptor). Although the dictionary size is different (150 for complete images and 70 for patches on edge maps), we have a great difference in AUC (90.3% and 87.4%).

Chapter 4

Advancing Bag-of-Visual-Words Representations for Lesion Classification in Retinal Images

This chapter presents an explanation about the methods of bag of visual words as well as a new technique developed for assignment, that shows be suitable for DR-related lesion detection. In this chapter, we extend upon a previous work of our group which explored a unified approach to detect bright (hard exudates) and red (microaneurysms and hemorrhages) lesions (See Fig. 4.1) [72]. We detect more lesions and substitute the quantization step, limited to hard-sum, for other alternatives (including this one) which will be explained herein. The methods developed herein resulted in a paper submitted to a top-tier journal currently under review.

4.1 Preamble

For progressive diseases, such as the many complications of *diabetes mellitus*, early diagnosis has a huge impact in prognosis, allowing corrective or palliative measures before irreversible organ damage takes place. In the case of Diabetic Retinopathy (DR), a common complication of diabetes mellitus, early detection is often crucial to the preservation of visual function. Therefore, screening patients for the characteristic lesions of DR is an important prophylactic measure. However, in poor, rural or isolated communities, the access to healthcare professionals — particularly to specialists — might be too precarious to ensure such prophylaxis.

In such scenarios, aided diagnosis may be very helpful. Eye-fundus images can be automatically processed in order to verify if the patient should be referred to an ophthalmologist for further investigation. However, in order to be useful, such systems must be

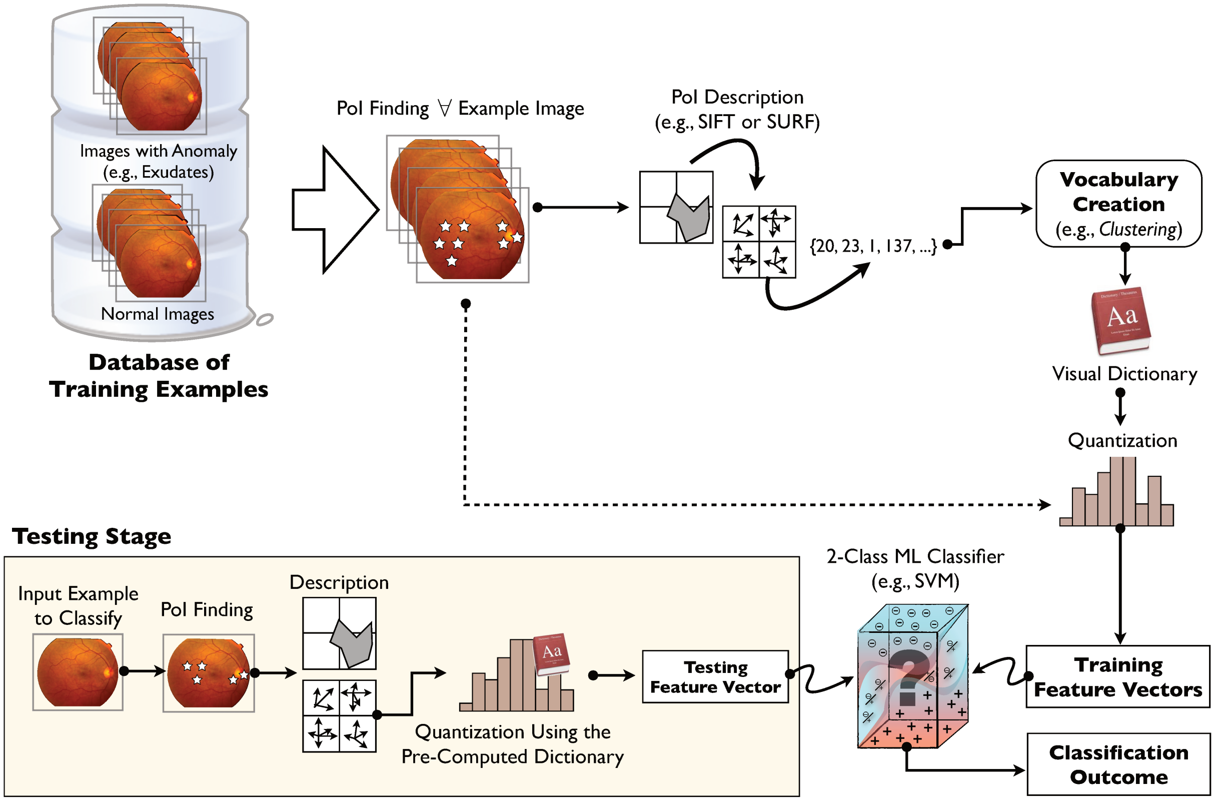

Images with Anomaly (e.g., Exudates)
Normal Images
Database of Training Examples
PoI Finding ∀ Example Image
PoI Description (e.g., SIFT or SURF)
{20, 23, 1, 137, ...}
Vocabulary Creation (e.g., Clustering)
Aa
Visual Dictionary
Quantization
Testing Stage
Input Example to Classify
PoI Finding
Description
Quantization Using the Pre-Computed Dictionary
Testing Feature Vector
2-Class ML Classifier (e.g., SVM)
Training Feature Vectors
Classification Outcome

Figure 4.1: Scheme of previous work of our group for the detection of bright and red lesion, exploiting only the hard-sum assignment technique [72].

accurate, because neither we want to leave a patient in need without care, nor we want to deluge the healthcare professionals with unneeded referrals (as we discuss in Chapter 6).

DR diseases manifest as different types of lesions, whose particularities are often employed in detection algorithms, involving the pre-processing of images and many *ad hoc* decisions [76, 39, 29, 30, 80, 88].

In our work, we employ a different strategy. We use a unified methodology, based on bag-of-visual-words (BoVW) representations, associated to maximum-margin support-vector machine (SVM) classifiers. Such methodology has been widely explored for general-purpose image classification, and consists of the following steps: (i) extraction of low-level local features from the image; (ii) learning of a codebook using a training set of images; (iii) creation of the mid-level (BoVW) representations for the images based on that codebook; (iv) learning of a classification model for one particular lesion, using an annotated training set; (v) using the BoVW representation and the learned classification model to make decisions on whether or not a particular image has a lesion.

The creation of mid-level BoVW representations can be further decomposed into two steps: the *coding* of the low-level feature vectors using the codebook, and the *pooling* of the codes obtained into a single aggregated feature vector [12]. There are several options for the coding and pooling operations. In this work, we go beyond prior work that have considered visual words for detecting DR-related lesions in eye-fundus images [72, 42, 40]. We explore several combinations of alternatives for the extraction of low-level features, and the creation of mid-level features pointing out important choices we might perform for boosting lesion detection in eye-fundus images.

Given the achievements that we shall detail in the next sections, we can anticipate the sparse technique associated with the semi-soft assignment represents an important breakthrough in comparison with the state-of-the-art, improving both the speed and accuracy of the methods specially regarding the detection of difficult lesions such as cotton-wool spots and drusen.

We organized the remainder of this chapter in four subsections. Section 4.2 presents the state of the art in two parts, one dedicated to DR-related lesion detection (Section 4.2.1) and one dedicated to the BoVW model (Section 4.2.2). In Section 4.3 we discuss the proposed scheme, starting with a discussion of the BoVW representation for DR-lesions (Section 4.3.1), an explanation of the proposed Semi-soft coding (Section 4.3.2) and the class-aware codebook creation (Section 4.3.3). The experiments are in Section 4.4, which starts with a detailed description of the datasets and protocols (Section 4.4.1) and finishes with the results (Section 4.4.2), including the statistical design, employed in the evaluations. Finally, in Section 4.5, we conclude the chapter and point out future research directions.

4.2 Related work

This section presents the state of the art dedicated to DR-related lesions and BoVW model.

4.2.1 Diabetic Retinopathy

Diabetes mellitus is a chronic end-organ disease that affects the circulatory system, including the retina, where it triggers Diabetic Retinopathy (DR). DR is the major cause of blindness in Europe and the U.S, in people of working age. It is a silent disease, whose symptoms often appear at late stages, when damage is already widespread [64].

According to the International Diabetes Federation[1], that prevalence may reach as much as 552 million people by the year 2030. Since the number of ophthalmologists is not growing at the same rate, there is a concern that medical personnel will be unable to cope

[1] `http://www.idf.org/diabetesatlas/5e/diabetes`

with the staggering amount of patients. Therefore, an automated and accurate screening tool can be, in the near future, an important adjunct in diabetes clinics, helping to refer to ophthalmology specialists only those patients in need of further attention [34, 29]. That may be particularly important for poor, isolated or rural communities, where the full-time presence of an ophthalmologist is unfeasible and costly.

The literature on aided diagnostics for DR tends to be specialized for each types of lesion [76, 39, 29, 30, 80, 88]. The results obtained are satisfactory for use as screening tools devoted for specific lesions. For instance, for white lesions detection, sensitivities range from 70.5 to 95.0% and specificities from 84.6 to 98.8% [30, 80, 88]; for red lesions detection, sensitivities range from 77.5 to 85.4% and specificities from 83.1 to 90.0% [76, 39, 29].

That specialization is a limitation found in many works: in general, a method developed for one lesion cannot be directly applied to detect another lesion, preventing the development of a general framework to detect any kind of DR-related lesion. Since there are several different DR-related lesions, a unified detection framework would be very desirable. It is worth noting, however, that some efforts are already being made towards this direction. Li et al. [46] have implemented a system for providing a management of diabetic eye disease in real time that focuses on the two major lesions associated with diabetes: microaneurysms and hard exudates. However, the framework does not exploit a unique technique for the detection of both lesions. After the detection process, the automated diagnosis is given by content-based image retrieval approaches.

Another common limitation of specific-lesion schemes is the need for complex and *ad hoc* pre- and post-processing of the retinal images, depending on the lesion of interest, and conditions of acquisition, fields of view and even ethnicity of the patients [34, 19]. The preprocessing, considering the analysis of retinal images, often includes resolution and color normalization, segmentation for detection of blood vessels, and detection and removal of the optical disk [29, 4]. Morphological operators [36] are often employed [30, 80, 88]. The post-processing step in eye-fundus images may include the identification of the retinopathy stage (mild, moderate and severe) based on the counting of the number of discontinuities, and the disposal of any response whether it does not attend a minimum criterion of reliability.

Sinthanayothin et al. [76] have developed a method for the detection of both bright and red lesions. They used several preprocessing techniques that begin with a conversion of the color space from RGB to IHS, contrast enhancement in the intensity band, and conversion back to the original color model. Thereafter, a recognition of the retinal elements according to the lesion of interest is performed. For the exudate detection (bright lesion), the authors performed a recursive region growing segmentation (RRGS) step for identifying similar pixels, which satisfy some criteria, such as gray level, within a region to determine the location of a boundary. The median intensity of the background (resulting region with more pixels) was set as a threshold to differentiate exudate from non-exudate pixels. The detection of hemorrhages and microaneurysms is performed in the green color channel as it contains more information and greater contrast for red lesions. In order to sharpen the edges against the red-orange background, the authors applied the Moat Operator technique. Then, they classified the retinal images using the same method for bright lesions, employing RRGS and thresholding.

The variety of methods for the detection of diabetic retinopathy is not restricted to the detection of specific lesions. Some methods aim at the identification of the current stage of the disease. Nayak et al. [52], for example, have used morphological operations and texture analysis for the extraction of features to be used as input for an automatic classification algorithm with neural networks. The features are related to the area of blood vessels, area of hard exudates and texture. The neural network classifies the images in one out of three classes: two for disease-related (non-proliferative retinopathy and proliferative retinopathy) and the normal class.

Yun et al. [95] have also used morphological operations and neural networks for the identification of the DR different stages. The process begins with contrast improvement, histogram equalization, morphological operations and binarization. After preprocessing the images by means of the morphological operations, the features are extracted counting the pixels contained in the perimeter and the area for each RGB layer, resulting in six features. Four groups are identified: normal retina, moderate non-proliferative retinopathy, severe non-proliferative retinopathy and proliferative retinopathy. The work developed by Nayak et al., as well as the paper of Yun et al. achieved a sensitivity of 90.0% and a specificity of 100.0%.

Recent works are becoming more general and bypassing the need of pre- and post-processing. Rocha et al. [72] have proposed a unified framework for detection of both hard exudates and microaneurysms. The authors have introduced the use of bags of visual words (BoVW) representations for DR-related lesion detection, creating a framework easily extendible to different types of retinal lesions. However, the BoVW model employed in that work is very simple and chosen without any theoretical or experimental design analyses opening the opportunity for substantial improvements, which we explore in this chaper. Furthermore, in this chapter, the evaluation of the alternative combinations for the BoVW is performed in a more statistically rigorous experimental design supporting all our claims and decisions.

We have also investigated fusion schemes for obtaining decisions from the evidences of specific anomaly detectors [40] (Chapter 5 presents the details of such work). The decision process for referring or not a patient from individual lesion classifiers has also been a topic of study. Niemeijer et al. [57] have combined different detectors for specific lesions into a single automatic decision scheme. Chapter 6 contains an advanced method employed in this work for the referral problem.

4.2.2 BoVW Representations

Representations based upon the aggregation of encoded local features have become a staple of the image classification literature. The technique has been definitely popularized by the work of Sivic and Zisserman [78], who have made explicit an analogy with the traditional bag-of-words representation used in Information Retrieval [8]. In their formalism, we reinterpret the local image descriptors as "visual words" by associating them to the elements of a codebook, which is aptly named a "visual dictionary". If we count the visual words for a given image, we obtain a histogram named bag-of-visual-words (BoVW), which is then used as a mid-level representation.

Learning the codebook is a challenge for BoVW representations. The traditional way involves unsupervised learning over a set of low-level features from a training set of images. K-means clustering, for example, can be used on a sample of those features and the k centroids be employed as codewords. There is also considerable variation throughout literature on the *size* of the codebook, ranging from a few hundred codewords until hundreds of thousands. Schemes could be employed to find the best size for each lesion detector [58].

The metaphor of "visual word" should not be taken too literally. While textual words are intrinsically semantic, visual words are usually appearance-based only. Moreover, the BoVW model has been considerably extended since the seminal work of Sivic and Zisserman. New ways of encoding the local descriptors using the codebook have been proposed, as well as new ways of aggregating the codes obtained. That stretches the metaphor of "visual word" too much, and a more formal model has been promoted by Boureau et al. [12], making explicit the operations of *coding* and *pooling*. Therefore, the BoVW formalism has evolved into a meta-model for which myriads of variations are possible, according to the combinations of low-level descriptor, codebook learning, coding and pooling.

The coding and pooling operations can be conveniently understood in the matrix form proposed by Precioso and Cord (see Fig. 4.2, adapted from [68, 7]). We suppose the codebook already given, as an indexed set of vectors, sampled or learned from the low-level feature space, $\mathcal{C} = \{\mathbf{c}_i\}$, $i \in \{1, \dots, M\}$, where $\mathbf{c}_i \in \mathbb{R}^d$. Then, for a given image, we start with the set of local descriptors $\mathcal{X} = \{\mathbf{x}_j\}$, $j \in \{1, \dots, N\}$, where $\mathbf{x}_j \in \mathbb{R}^d$ is a local feature and N is the number of salient regions, points of interest, or points in a dense sampling grid. We call $\mathbf{z}$ the final BoVW vector representation [12, 7].

The coding step transforms the low-level descriptors into a representation based upon

$$\mathbf{H} = \begin{array}{cc}
 & \begin{array}{ccccc} \mathbf{x}_1 & \cdots & \mathbf{x}_j & \cdots & \mathbf{x}_N \end{array} \\
\begin{array}{c} \mathbf{c}_1 \\ \vdots \\ \mathbf{c}_m \\ \vdots \\ \mathbf{c}_M \end{array} & \left[\begin{array}{ccccc}
\alpha_{1,1} & \cdots & \alpha_{1,j} & \cdots & \alpha_{1,N} \\
\vdots & & \vdots & & \vdots \\
\alpha_{m,1} & \cdots & \alpha_{m,j} & \cdots & \alpha_{m,N} \\
\vdots & & \vdots & & \vdots \\
\alpha_{M,1} & \cdots & \alpha_{M,j} & \cdots & \alpha_{M,N}
\end{array}\right] \Rightarrow g : \mathit{pooling} \\
 & \Downarrow \\
 & f : \mathit{coding}
\end{array}$$

Figure 4.2: The BoVW model illustrated in a convenient matrix form, highlighting the relationships between the low-level features $\mathbf{x}_j$, the codewords $\mathbf{c}_m$ of the visual dictionary, the encoded features α_m, the coding function f and the pooling function g.

the codewords, hopefully one better adapted to the specific task, one that preserves all relevant information, while discarding noise. Coding can be modeled by a function f: $\mathbb{R}^d \rightarrow \mathbb{R}^M$, $f(\mathbf{x}_j) = \alpha_j$ that takes the individual local descriptors $\mathbf{x}_j$ and maps them onto individual codes α_j. The classical BoVW model employs the "hard assignment" of a low-level descriptor to the closest codeword, and can be modeled by:

$$\alpha_{m,j} = 1 \text{ if } m = \arg\min_k \|\mathbf{c}_k - \mathbf{x}_j\|_2^2 \quad \text{else} \quad 0 \tag{4.1}$$

where $\alpha_{m,j}$ is the m^{th} component of the encoded descriptor.

Recent literature [12, 85], however, suggests that "soft" coding schemes, which allow degrees of association between the low-level descriptors and the elements of the codebook work better, avoiding both the boundary effects and the imprecisions of hard assignment [85].

The pooling step takes place after the coding, and can also be represented by a function g: $\{\alpha_j\}_{j\in 1,\ldots,N} \rightarrow \mathbb{R}^M$, $g(\{\alpha_j\}) = \mathbf{z}$. The classical BoVW corresponds to a "counting of words" (called sum-pooling) and can be modeled as:

$$g(\{\alpha_j\}) = \mathbf{z} : \ \forall m, \ z_m = \sum_{j=1}^{N} \alpha_{m,j} \tag{4.2}$$

That simplistic pooling has also been criticized, and taking the maximum activation of each codeword (in a scheme aptly named max-pooling) is often much more effective [13]:

$$g(\{\alpha_j\}) = \mathbf{z} : \ \forall m, \ z_m = \max_{j\in\{1,\ldots,N\}} \alpha_{m,j} \tag{4.3}$$

The vector $\mathbf{z} \in \mathbb{R}^M$ obtained from pooling is the BoVW representation, used for classification. Those vectors are often normalized: for example, in the classical BoVW scheme, ℓ_1-normalization is often employed to turn a vector of occurrences into a vector of relative frequencies.

4.3 Proposed Methodology

In this section, we present in detail the proposed scheme for DR-related lesions detection, as well as the alternatives considered, and which will be evaluated in Section 4.4. The scheme is based upon two steps (i) a mid-level BoVW-based representation; and (ii) a maximum-margin SVM classification.

For the classification model, we have employed a maximum-margin Support Vector Machine (SVM) [18] with Gaussian kernel as the final BoVW classifier. The classifier parameters C (the margin "hardness", an inverse regularization parameter) and γ (the standard deviation of the kernel) were found by cross-validation, using the standard Lib-SVM's built-in grid-search fine-tuning algorithm [16].

The representation, which is the main contribution of this chapter, is dissected in the next three subsections.

4.3.1 BoVW-based representation

As seen in Section 4.2.2, a BoVW-based representation rests upon several choices. In this section, we explain in detail our particular implementation, as well as the variations we have considered in our experiments for detecting different DR-related lesions in eye-fundus images.

The factors we consider here are:

- **Low-level feature extraction**: the mid-level BoVW features are based upon the low-level features, whose choice has great impact on performance. Two treatments are usual: **sparse features**, based upon the detection of salient regions, or points-of-interest; and **dense features**, sampled over dense grids of different scales;

- **Choice of codebook**: a challenging step, the codebook learning is usually performed by a k-means clustering over features chosen **at random** from a training set of images. We evaluate an alternative **class-aware** treatment, described in Section 4.3.3;

- **Coding**: besides the traditional **hard** assignment, we have tested one of the **soft** assignments proposed by Gemert et al. [85], and a new **semi-soft** assignment, especially conceived for the DR-related lesion detection application, described in Section 4.3.2.

We highlight that BoVW-based representations have already been proposed in the literature [72, 40, 65]. However, the methods discussed in those papers do not stretch out the several choices associated with BoVW-based representations nor do they present any elaborate discussion on the rationale for using the representations proposed therein.

For the low-level feature extraction, we employ SURF descriptors [10]. Compared to the obvious alternative, SIFT [48], SURF has shown superior results in previous evaluations on DR-related lesion detection [72, 40, 65], besides being faster. SURF takes advantage of the integral images technique to allow the accelerated computation of rectangular convolution masks. The salient-region detector is based upon an integer approximation of the Hessian matrix, while the extracted descriptors are based upon sums of 2D Haar wavelet responses, both of which can be computed extremely fast with integral images.

For the *sparse features* treatment we employ both the detector and descriptor, as implemented in SURF version 1.0.9, released by Bay et al. [10]. For the *dense features* treatment, we employ only the descriptor, densely sampled on grids of scales of radii 12, 19, 31, 50, 80, 128 pixels, corresponding approximately to the scales of interest of the lesions and groups of lesions, measured by hand in a sample of the images, that is, using a few selected images with lesions, we analyzed them directly and measured the average number of pixels related to the lesions. Those scales correspond both to the σ of the Gaussian window of SURF, and to the sampling step of the dense grid. Therefore, there is about 50% overlap between consecutive samples in the grid.

SURF has some sensitivity parameters, that we tuned so that 400 (empirically defined) points of interest (PoIs) are detected on average per image, after changing the threshold (different for each dataset) and filtering the points over the external edges of the retinas. We also adjusted the parameters to get an extended descriptor with 128 dimensions and to operate using the double image resolution. These modifications are not considered pre-processing steps since they are common to all images and are made simply to ensure a minimum number of points of interest per analyzed image.

For the codebook learning, in the first step, we select the candidate local descriptors within the regions of interest marked by a medical specialist in the set of training images. We have a set of 19,170 candidate points for normal images and an average of 2,820 points within regions marked as having lesions. Then, we employ k-means clustering for $k = 250$ in two turns, one for each class, resulting in a set of 500 centroids (a value which was known to work well from previous works of our group [72]). We allow k-means to run for at most 200 rounds (each stage represents the choice of new candidate centroids and their associated distortion) or until convergence (the distortion for different groups are small), which comes first. The 500 cluster centroids are then used as codewords. The *class-aware* treatment is explained in detail in Section 4.3.3.

For most comparisons we perform in this chapter, we set the size of the codebook to 500 codewords, but due to the counter-intuitive result that the dense extraction performed worse than the sparse, we have explored a larger dictionary of 1,500 words for that treatment, in order to evaluate if there was a correlation between the extraction of more features and the need for larger samples in the dictionary.

For the coding step, we test three treatments:

- **Hard assignment**: associates each descriptor fully and only to its closest codeword in the visual dictionary, as explained in equation 4.1. The advantage of those schemes is the sparsity of the codes; the disadvantages, already mentioned, are that they are subjected to imprecisions and noise, when the descriptors fall in regions close to the limit between the codewords in the feature space. This scheme was explored in previous work for detecting DR-related lesions in eye-fundus images [72, 42, 40].

- **Soft assignment**: there are several "soft" schemes, all trying to cope with the deficiencies associated with the hard assignment treatment. The one we employ is called *codeword uncertainty* [85] and is generally considered the most effective:
$$\alpha_{m,j} = \frac{K_\sigma(\|\mathbf{c}_m - \mathbf{x}_j\|\|_2)}{\sum_{\mathbf{c}\in\mathcal{C}} K_\sigma(\|\mathbf{c} - \mathbf{x}_j\|\|_2)}, \tag{4.4}$$
where K_σ is the Gaussian kernel. We employ $\sigma = 45$, a value derived observing a population of distances between pairs of SURF descriptors in a very large dataset of images. This treatment was never explored in the DR-related lesion detector literature.

- **Semi-soft assignment**: soft assignment solves the boundary effects of hard assignment, but creates codes which are too dense. A "semi-soft" scheme is often more desirable. One such scheme, designed specially for the DR-related lesion detection, is described in Section 4.3.2.

For the pooling step, we forgo the traditional **sum**-pooling (Eq. 4.2), and employ the more recent **max**-pooling, described in Eq. 4.3). The pooling step is considered one of the most critical for the performance of BoVW representations, and max-pooling is considered an effective choice [12, 13, 7].

In all cases, we employ an ℓ_1-normalization in the final BoVW vector.

4.3.2 Semi-soft coding

The semi-soft coding tries to combine the advantages of both hard and soft assignments, i.e., avoiding the boundary effects of the former, and the dense codes of the latter. The main idea is to perform a soft assignment, but just to the few codewords which are the closest to the descriptor, keeping all others at zero. That general idea can be translated into many designs. The one we propose here is based upon two simple principles:

- only the closest codeword is activated;
- the activation is proportional to the inverse of the distance between the codeword and the descriptor.

Therefore, the generated codes are very sparse. On the other hand, the effect of the descriptors is "felt" even at relatively long distances (compared to exponential, or power-law decays). The scheme has the advantage of requiring no parameters.

The coding function can be described as:

$$\alpha_{m,j} = \begin{cases} \frac{1}{\|\mathbf{c}_m - \mathbf{x}_j\|_2} & \text{if } m = \arg\min_k \|\mathbf{c}_k - \mathbf{x}_j\|_2 \\ 0 & \text{otherwise,} \end{cases} \tag{4.5}$$

4.3.3 Class-aware codebook

Rocha et al. [72] have proposed employing a "double codebook", extending the usual scheme in a class-aware fashion, especially adapted for DR-related lesions. That is possible because, in addition to the training images being annotated for each lesion, also the regions where the lesions appear are identified (usually two to five per image from affected patients).

The idea of using the class-aware codebook is to ensure that the appearances characteristic of the lesions are well-represented during the coding phase, instead of counting on luck alone. Selection of feature vectors is usually employed for general-purpose visual recognition – but in those tasks, recognition does not hinge in such subtle differences as is the case for DR-related lesions. The scheme can be employed for both dense and sparse low-level descriptors, and is illustrated for the latter case in Fig. 4.3.

The class-aware scheme works by creating two independent codebooks, one from descriptors sampled from regions marked as containing lesions, and one from descriptors outside those regions (which includes images from healthy patients). Then, two independent k-means clusterings are performed, each with k corresponding to half the size of the desired codebook. After the clustering is finished, the two sets of centroids are simply concatenated, generating a codebook of the desired size.

As we cited before for the codebooks creation, the k-means procedure was executed in at most 200 rounds or until convergence, one time for each class.

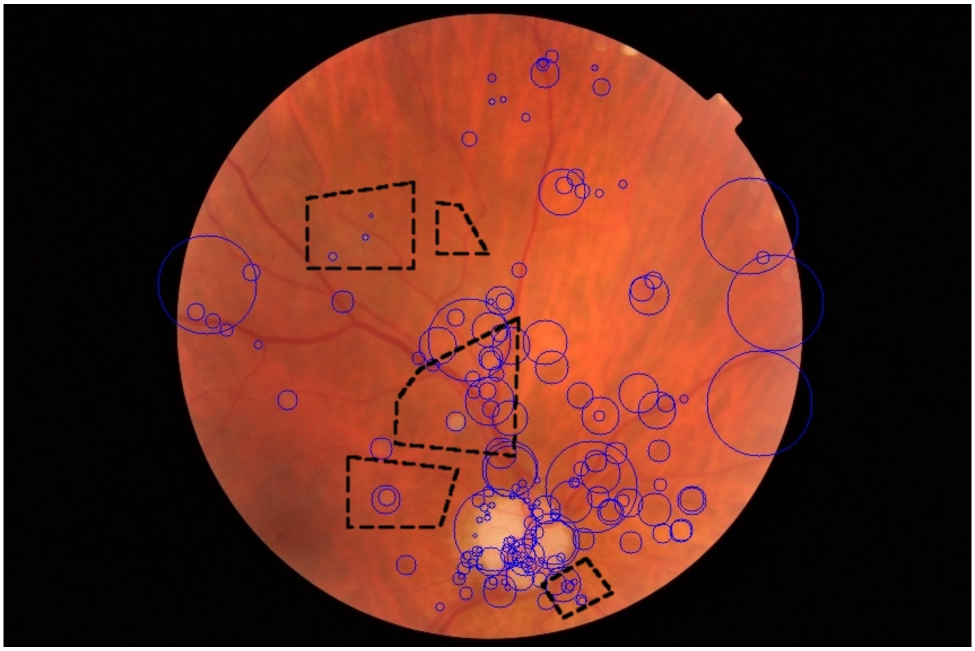


Figure 4.3: Regions of interest marked by a medical specialist (dashed black regions) and the points of interest extracted in the sparse technique (blue circles). Points of interest falling within the regions marked by the specialist are further considered for creating the BoVW representation of a lesion while points found in normal images are used for the BoVW representation of images of healthy patients. In the class-aware codebook, both representation are combined.

4.4 Experiments

4.4.1 Data, protocol and metrics

We performed the experiments using three different retinal image datasets annotated by medical specialists:

- **DR1 dataset**, provided by the Department of Ophthalmology, Federal University of São Paulo (Unifesp). Each image was manually annotated by three medical specialists and all the images in which the three annotations agree were kept in the final dataset. The images were captured using a TRC-50X (Topcon Inc., Tokyo, Japan) mydriatic camera with maximum resolution of one megapixel (640 × 480 pixels) and a field of view (FOV) of 45 degrees.

- **DR2 dataset**, from the same source, after discarding the poor quality images [65]. The dataset was captured using a TRC-NW8 retinograph with a Nikon D90 camera, creating 12.2 megapixel images, which were then reduced to 867 × 575 pixels for accelerating computation.

- **Messidor dataset**, captured in three different French ophthalmologic departments. There are three subsets, one for each department. The images were captured using a Topcon TRC-NW6 non-mydriatic retinograph with a 45 degrees field of view, at the resolutions of $1,440 \times 960$, $2,240 \times 1,488$ or $2,304 \times 1,536$ pixels.

Both DR1 and DR2 datasets are publicly available[2]. The Messidor dataset is also available to the scientific community, after a registration is fulfilled[3]. Statistics about the three datasets are given in Table 4.1.

All experiments were performed using a cross-dataset protocol, an important precaution in the design, since in clinical practice the images that need to be classified seldom will be acquired in the exact same conditions (camera, resolution, operator, FOV) than the images used for training. We emphasize that the datasets were collected in very different environments with different cameras, at least one year apart and in different hospitals. We have employed the entire DR1 as the training dataset. The DR2 and Messidor were then employed for testing.

The cross-dataset protocol poses experimental design challenges, because of the different standards used in the annotations of the three datasets. In DR1, images are annotated with the specific tags *deep* and *superficial hemorrhage*. In DR2, only the general *red lesion* tag is employed. In Messidor, the images are annotated not only for the presence

[2] `http://www.recod.ic.unicamp.br/site/asdr`

[3] `http://messidor.crihan.fr`

Table 4.1: Annotation occurrences for the three datasets

Lesion	DR1	DR2	Messidor
Hard Exudates (HE)	234	79	654
Superficial Hemorrhages (SH)	102	—	—
Deep Hemorrhages (DH)	146	—	—
Red Lesions (RL)*	—	98	226
Cotton-wool Spots (CS)	73	17	—
Drusen (D)	139	50	—
Other lesions, excluding above	—	71	—
All lesions**	482	149	654
Normal (no lesions)	595	300	546
All images	1,077	520	1,200

* "Red Lesion" is a more general annotation that encompasses both SH and DH, besides microaneurysms.

** The lesions do not sum to this value because an image can present different types of lesion at once.

of the lesions, but also for the severity, evaluating the number of microaneurysms and hemorrhages (red lesions), the presence or absence of neovascularization (not evaluated in this work), and the proximity of the exhudates to the macula. In order to make the cross-dataset classification possible, and the joint statistical analysis of the two sets of experiments (DR2 and Messidor) feasible equivalences were found between the datasets, as detailed in Table 4.2.

Table 4.2: Composition of the cross-dataset training and test

	Train	Test	
Lesion	DR1	DR2	Messidor
Hard Exudates (HE)	234	79	654
Superficial Hemorrhages (SH)	102	—	—
Deep Hemorrhages (DH)	146	—	—
Red Lesions (RL)*	180	98	226
Cotton-wool Spots (CS)	73	17	—
Drusen (D)	139	50	—

* The annotations SH and DH are added to form the training set in DR1, summing 180 images due to the overlap.

To allow quantifying precisely the performance of the proposed method and enabling reliable comparisons, we employ receiver operating characteristic curves (ROCs), which plot the compromise between specificity (few false positives) and sensitivity (few false negatives). Whenever we needed to quantify performance as a single scalar, we have employed the area under the ROC curve (AUC). Since the classifier can trade specificity for sensitivity, the AUC gives a better overall performance measure than any particular point of those two metrics.

4.4.2 Results

The detailed results are presented in Tables 4.3 and 4.4, which show, respectively for the DR2 and Messidor datasets, the AUCs obtained for each lesion.

Row-by-row results of Tables 4.3 and 4.4 suggest the best configuration of the BoVW for each lesion (and dataset): the results tend to favor the semi-soft coding on sparse features, except for the drusen, which tend to favor dense features. The Messidor dataset, which has some very challenging images (patients with very early DR signs, showing very few lesions) also tends to favor dense features, but works well under the sparse features / semi-soft coding scheme.

However, such local, case-by-case analysis, fails to account for random effects. A less naïve analysis must take into account all results across BoVW parameters, datasets and lesions. Our goal is to obtain an overall best configuration for the BoVW, if such configuration can be found with confidence.

To perform the global analysis, we run a factorial analysis of variance (ANOVA) on the following factors (and levels):

(1) low-level feature extractor (Sparse, Dense),

(2) density of the sparse extractor/dictionary size of the dense extractor (Low, High),

(3) coding (Soft, Semisoft, Hard), with repeated measures for each Lesion (HE, RL, CS, D), and

(4) test dataset (DR2, Messidor).

All errors are measured within-subjects. Unfortunately, we immediately face an obstacle brought by the different annotation standards between DR2 and Messidor: the former has annotations for all four levels of lesion, but the latter only has annotations for hard exudates (HE) and red lesions (RL). Because it is challenging to perform (and to interpret) such unbalanced experimental designs, we have decided to perform two separate balanced studies: one considering only DR2 and all four lesions; another for both

Table 4.3: AUCs in %, for Training with DR1, Testing with DR2

	Sparse features			Dense features		
	Hard	Semi-soft	Soft	Hard	Semi-soft	Soft
Hard Exhudates (HE)	93.1	**97.8**	95.5	94.5	95.6	95.6
Red Lesions (RL)	92.3	**93.5**	87.1	89.1	90.6	89.9
Cotton-wool Spots (CS)	82.1	**90.8**	84.9	84.5	90.4	90.3
Drusen (D)	66.5	82.8	62.6	**84.1**	82.5	75.5

Table 4.4: AUCs in %, for Training with DR1, Testing with Messidor

	Sparse features			Dense features		
	Hard	Semi-soft	Soft	Hard	Semi-soft	Soft
Hard Exhudates (HE)	64.4	70.3	66.2	**70.5**	70.0	70.0
Red Lesions (RL)	77.4	83.1	76.6	**85.2**	85.1	82.5

test sets, but only HE and RL lesions. To remove the strong scaling effect of the lesions and datasets, we independently standardize each subject (lesion, dataset combination), subtracting its average AUC and dividing by its standard deviation.

The analysis on the DR2 subset finds an important interaction effect: the combination between the choice of Low-level Features and Coding ($p = 0.007$). The main effect of Coding alone just fails significance ($p = 0.062$), and all other effects and interactions are non-significant. An examination of Table 4.3 reveals why the factors are significant only in interaction, since the two low-level feature extractors seem to work better with different coding schemes. The synergy between sparse feature extraction and semi-soft coding for DR-lesion classification can be better appreciated in the box-plot of Fig. 4.4, that shows the within-subjects standardized AUCs for the six combinations of feature extraction and coding. The analysis on the subset with both test datasets and only HE and RL lesions shows similar results, with significant interaction between low-level feature extraction and semi-soft coding ($p = 0.011$).

A comparison with Rocha et al.'s paper [72], in which the class-aware scheme is proposed for the detection of bright and red lesions exploiting the classical hard-sum approach, makes it evident that our technique proposed for feature extraction may be suitable for DR-related lesion detection. The authors reached AUCs of 95.3% and 93.3% respectively for bright and red lesions, while our respective results are 97.8% and 93.5% for testing with DR2. The authors do not present detectors for additional lesions as we do in this work for cotton-wool spots (AUC = 90.8%) and drusen (AUC = 82.8%).

A crucial factor which has to be noted is the validation protocol. We performed the training and testing using distinct datasets, exploring the cross-validation protocol which is more robust than the 5-folds cross-validation used by the authors. Despite using

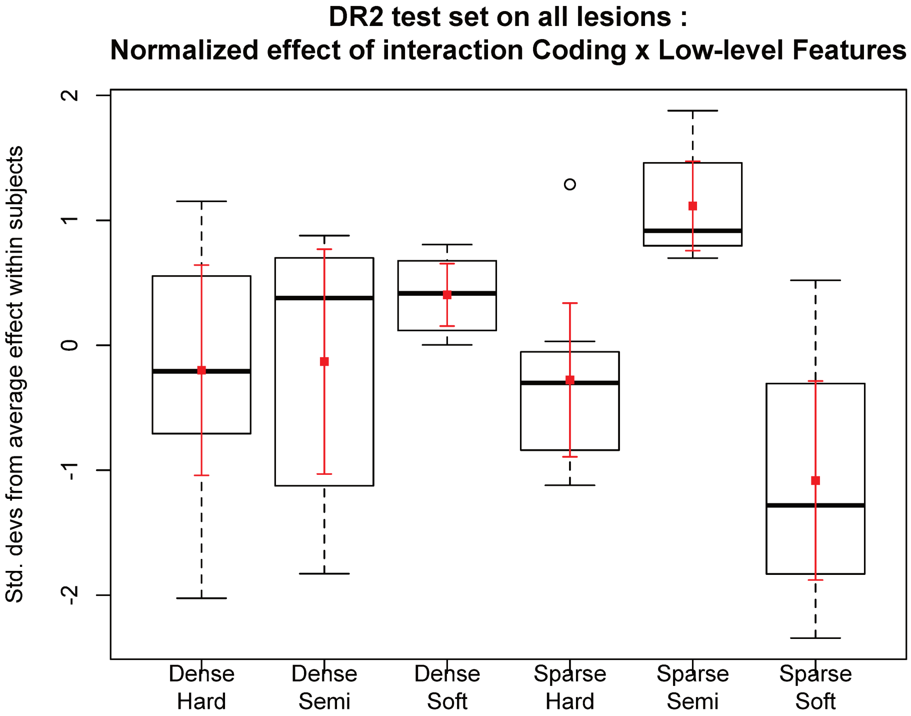


Figure 4.4: Box-plot for the within-subjects (per lesion) standardized AUCs for six combinations of feature extraction and coding, with averages and 95%-confidence intervals (in red), the whiskers show the range up to 1.5× the interquartile range, and outliers are shown as small circles. The strong synergy between sparse feature extraction and semi-soft coding is evident. This plot is based on a balanced design with the DR2 dataset and all lesions, another balanced design with both datasets and two lesions show similar results.

different datasets for training and testing, our results for hard-sum are lower but well-comparable to the authors ones (93.1% for bright lesions and 92.3% for red lesions).

4.5 Final Remarks

Automatic lesion detection systems have been crucial tools for facilitating attendance in rural and remote communities and for providing a screening able to determine whether there is need for consultation, helping the work of the medical specialists. Moreover, by providing a detection in the early stages of the disease, these tools trigger a reduction in treatment costs.

This chapter focused on the detection of Diabetic Retinopathy related lesions. Several studies have obtained satisfactory results for the detection of DR-related lesions in the literature. However, in previous works, the detection of different anomalies normally relied upon the use of distinct approaches based on specific properties of each lesion. This renders the detection of DR an expensive method, since it requires the execution of multiple detection procedures each one with specific parameters that need to be set up and learned. On the other hand, recent advances in DR-related lesion detection using approaches based on bags of visual words addressed the need of pre- and post-processing operations. However, such approaches employed some techniques without any theoretical or experimental design analyses opening several opportunities for contributions and advances.

In this chapter, we explored recent advances regarding bags of visual words related literature including its formalization and stretched out possible combinations we might perform for detecting DR-related lesions in eye-fundus images. We explored several combinations of alternatives for the extraction of low-level features, and the creation of mid-level representations pointing out important choices when designing a unified framework for detecting DR lesions.

Our main contribution in this work is the proposal of a new coding scheme called semi-soft, which explores the advantages of the most traditional hard sum coding (sparse coding) as used in prior work for DR lesion detection [72] and soft assignments (which better deal with imprecisions and noise). As we showed in the experiments with a detailed experimental design evaluation through ANOVA, the semi-soft coding associated with sparse feature extraction provides a good balance for designing an efficient and effective DR-related lesion detector with results that outperform the ones in the literature. In addition, the proposed combination also provides excellent results for two hard-to-detect DR lesions: cotton-wool spots and drusen.

At least for the particular problem of DR-related lesion detection, the sparse feature extraction + semi-soft coding combination defies the status-quo established by the Computer Vision literature for general object recognition problems in which it is stated that soft assignment + dense sampling is the way to go.

The discovery of the best method that showed to be very effective for detection of DR-related lesions opens the opportunity for deploying the sparse technique with semi-soft coding to other applications. A possible future work consists of identifying the precise location of the lesion, as well as the size and quantity, and defining the DR severity degree of a patient further classifying the images as related to DR cases in early, mild, proliferative and severe stages.

Chapter 5

Data Fusion for Multi-lesion Diabetic Retinopathy Detection

This chapter contains the explanation of the methods used in this work for fusion, as well as outcomes achieved for multi-lesion detection. The methods developed herein resulted in the publication [40] and also in the paper submitted to a top-tier journal, detailed in the Chapter 4).

5.1 Preamble

In health care, the early diagnosis of disease has been important for maintaining optimal health and reducing costs associated with treatment, and has contributed to improve the patients' quality of life. Diabetic retinopathy, if not discovered and treated in time, can lead to the complete loss of sight. Identifying DR early through systematic screening and implementing timely treatment are important steps to prevent blindness [51].

Developing a unified framework that can identify different retinal lesions has been published using a visual words dictionary model [42]. However, this model creates a large set of visual words, which increases with the number of lesions that are identified. Therefore the lesion detectors need to be combined to optimize the classification. Detector fusion has been applied in areas such as face and object detection [27]. Specifically within the field of multi-lesion detection associated with DR, some methods have been applied [4] but require further development for the DR model.

The most common fusion methods can be classified into three levels: (1) abstract (each classifier outputs the class label for each input pattern); (2) rank (each classifier outputs a ranking list of possible classes for each input pattern); and (3) measurement (each classifier outputs a score, probability or confidence level for each input pattern) [93]. Among the abstract fusion methods, majority voting is the most discussed. On the measurement level ‘sum’, ‘product’, ‘max’, ‘min’, ‘average’, ‘median’, ‘OR’ and ‘AND’ methods [67] are commonly employed. Fusion methods on the rank level, such as Borda Count, may not be suited for classifying DR.

The current chapter presents a visual words framework that is able to identify DR-related lesions based upon the identification of the most common ones: hard exudates, deep hemorrhages, superficial hemorrhages, drusen, and cotton wool spots. Acharya et al. [4] reported multi-lesion detection in DR using mathematical morphology and support vector machine classification to detect exudates, hemorrhages, and microaneurysms.

Our main approach expands upon this and our previous work [40] and consists in investigating fusion of different detectors to identify the presence of DR. Points of interest are combined into visual words and a visual dictionary [78] that is able to identify specific anomalies within the retina is created.

Our work contains a set of classifiers that act in cooperation to solve a pattern recognition problem [43, 37], followed by several methods for classifier fusion. This kind of approach is intuitive since it imitates our nature to seek several opinions before making a crucial decision [73]. Two fusion methods were evaluated: OR and meta-classification.

Section 5.2 presents related work for automatic DR-related lesion detection based on classifier combination. Section 5.3 introduces our method based on classifier fusion for detecting different retinal pathologies. Section 5.4 reports the experiments and results. Finally, Section 5.5 presents final remarks.

5.2 Related work

Combining multiple classifiers is a standard practice in medicine. Examples of use of fusion methods are breast cancer prediction [71] and lung cancer detection from computed tomography scans [92].

Dimou et al. [24] evaluated the use of an ensemble of eight classifiers based on 15 different fusion strategies to provide accurate diagnosis of different types of cancer based on the available predictors. The authors demonstrated that a variety of different classifier fusion techniques can be used to augment the diagnostic performance of individual models in the context of practical biomedical applications.

There are not many reports that focus particularly on the problem of fusion methods for DR detectors. Niemeijer et al. [55] applied classifier fusion methods to combine several detectors of microaneurysms in retinal images. The authors used two basic approaches for the fusion: static combination rules (sum, product, and maximum) and meta-classification. The results indicated, in some cases, that combining detectors for the same lesion does not necessarily result in better performance compared to the best individual detector. However, in our case, we propose the fusion of classifiers specialized in specific but most common lesions, aiming the development of a robust screening framework.

5.3 Proposed Methodology

We propose to apply detector fusion to a multi-lesion detector algorithm based on a visual words dictionary [78]. This method is characterized by the approach in which detectors operate in parallel and are combined to obtain a result related to the presence of any DR-related lesion.

Visual dictionaries, which were detailed in Chapter 4, constitute a robust representation approach in which each image is treated as a collection of regions. In this representation, the only important information is the appearance of each region [89, 6].

The objective when creating a visual dictionary is to learn, from a training set of examples, the generative model that selects the most representative regions for a given problem. The number of selected regions must be large enough to distinguish relevant changes in the images, but not so large as to distinguish irrelevant variations such as noise [15].

5.3.1 Fusion of detectors

According to [90], there is not a single classifier that can be considered optimal for all problems. There are also no clear guidelines for choosing a set of machine learning methods for a specific task and it is rare to have complete knowledge of the data distribution and details of how the classification algorithm behaves. Therefore, it becomes difficult to classify a retinal image according to the presence of DR with a single method. Certainly, it is difficult or impossible to find a good single classifier trained to detect only one lesion but able to detect any evidence of diabetic retinopathy. Therefore multiple detectors have to be implemented. Currently, it is common that different pre- and post-processing procedures are required for each detector, making multi-lesion detection difficult and not very accurate [57, 1].

For this step, we have a set of detectors for six individual DR-related lesions: hard exudates, superficial hemorrhages, deep hemorrhages, cotton wool spots, drusen and red lesions (superficial hemorrhages or deep hemorrhages). The assignment approach explored for the development of the detectors is the semi-soft, explained in Chapter 4.

After creating a set of detectors, there are numerous methods for combining classifiers. The principal approach for combining classifiers is classifier fusion, which considers that all classifiers contribute to the final decision, assuming competitive classifiers [44]. In this work, we investigated two classifier fusion methods: OR and meta-classification.

OR

Included in the commonly used ensemble strategies, the logic OR is a fusion method of parallel architecture that labels as positive the data classified as positive in at least one classifier. Consequently, data is labeled as negative only if all the classifiers label it as negative. This method tends to obtain high sensitivity and low specificity.

Meta-classification

Meta-learning, employed in Chapter 2 for quality assessment, can be loosely defined as learning from information generated by different learners. In our work, we concentrate on learning from the output of inductive learning systems such as the SVM [11]. The output is defined as decision score. Meta-classification, in this case, means learning from the classifiers produced by the learners and the predictions of these classifiers on training data. A classifier (or concept) is the output of an inductive learning system and a prediction (or classification) is the predicted class generated by a classifier when an instance is supplied. Moreover, the training data presented to the learners initially are also available to the meta-classifier if warranted [15].

For obtaining the decision score d_i^m for a particular feature vector representing an image for one detector, we calculate the distance of feature vector representing such image to the decision hyperplane representing the detector. Fig. 5.1 depicts an example considering a linear classifier. This particular example shows a 2-D feature space.

The scores given by each classifier feeds an SVM as features. In this strategy, we are actually using a two-level classification with individual classifiers at the first level and a higher-level meta-classifier to learn over the individual classifiers combination.

5.4 Experimental results

In this section, we present the experiments we have performed to validate our approach.

For this chapter, we use the individual DR-related lesion detector which provided better results as showed in Chapter 4. As we concluded, the novel semi-soft coding/pooling BoVW representation highlighted over the other.

Given the implemented DR-related lesion detectors, and considering the method of sparse feature extraction with semi-soft assignment, here we present details about the technique employed for the development of a final detector whose objective is to point out whether an image is normal or has any lesion including possible ones not present during training.

As we cited before, the individual detectors were trained previously (see Chapter 4) using DR1 dataset. Given the detector models obtained, we test and extract a high-level

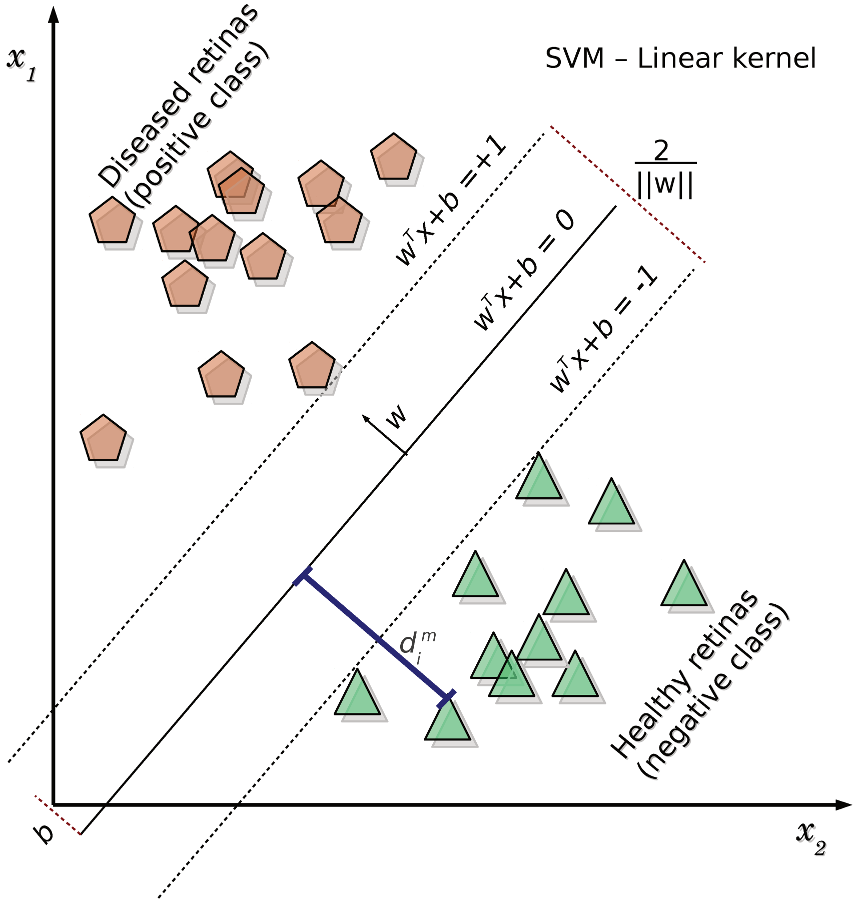


Figure 5.1: Example of an SVM classifier with linear kernel (lesion vs. normal).

description for DR2 dataset. Please refer to Section 4.4.1 for more information about the datasets.

We combined the results of each of the detectors using the logic OR technique. In principle, the combination is simple: if any detector identifies a lesion, the patient should be referred to the specialist. Here, we performed the fusion in three different testing steps:

(1) considering only images from DR2 for testing with at least one of the discussed anomalies (hard exudates, superficial hemorrhages, deep hemorrhages, cotton wool spots, drusen or red lesions);

(2) considering images from DR2 for testing with any DR-related lesion (including neovascularization, increased vascular tortuosity, foveal atrophy and chorioretinitis scar, for example); and

(3) considering images from DR2 for testing which present signals of other anomalies

(except the ones we trained for).

The objective of including other DR-related lesions in the testing set is evaluating the ability of the individual classifiers, even using a trivial fusion method, for detecting any lesion when operating in parallel and combined.

Figure 5.2 depicts the results achieved using the logic OR fusion for the considered anomalies. For the test with the same lesions (detecting anomalies already seen during training), the method achieved an AUC of 88.6%. When detecting all lesions (including possible ones not seen during training), the method yielded an AUC of 81.6%. Finally, evaluating the potential of the resulting classifier for detecting distinct lesions (not a single one it was trained with), the method obtained an AUC of 66.8%.

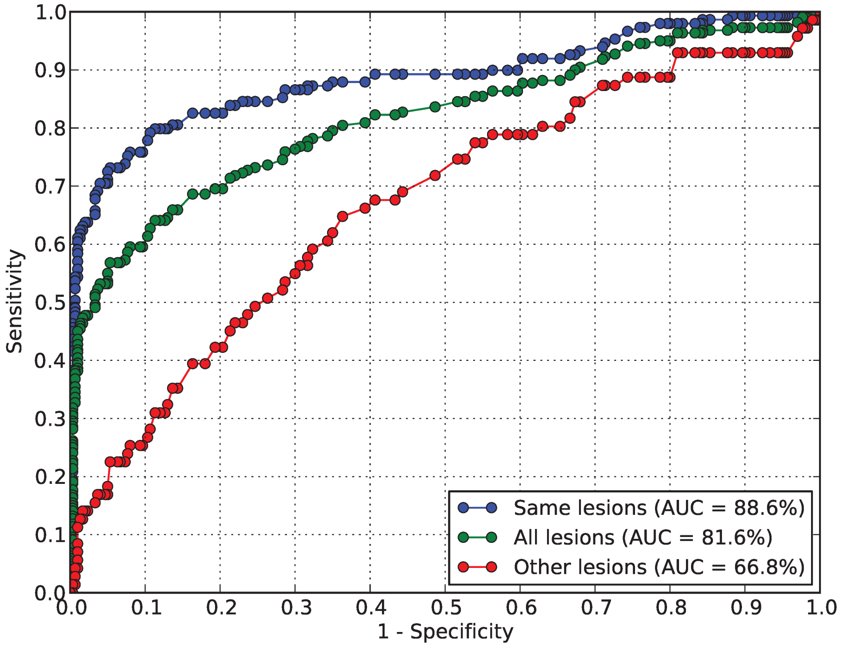


Figure 5.2: Results considering the logic OR fusion for the considered anomalies in three different scenarios.

The second approach explored for combining individual detectors in order to get a response about the presence of DR-related lesions is the fusion by meta-classification, which consists of using the outputs of a series of individual lesion classifiers as input to a new classifier at a higher level.

Note that in order to design the meta-classifier for deciding how to combine the DR-related lesion detectors outputs, we need to train it with suitable classifier outcome examples. For that, first we used the lesion detectors trained with DR1 using the method of sparse feature extraction with semi-soft assignment. Then consider a 5 × 2-fold cross-validation protocol [23] regarding DR2 dataset in order to have suitable examples for training and testing on such dataset. In each round out of five of this protocol, we divide DR2 into two sets and use one for training the meta-classifier and the other one for testing. Then we switch the sets. In the end, we have a reliable measure of the classifier accuracy and its variation for different testing sets.

Figure 5.3 depicts the ROC curves which expresses the mean and the standard deviation obtained for the meta-classification approach using the 5 × 2-fold cross-validation protocol. The area under the curve yielded by the meta-classification approach is equal to 89.3%±2.6% for same lesions. This fusion method, which have been explored in our previous work for the detection of DR-related lesions [40], outperforms the one obtained with the logical OR fusion technique (AUC = 88.6%). For the detection of any DR-related lesion, the meta-classification (AUC = 82.5%±2.5%) also outperforms the logical OR method (AUC = 81.6%).

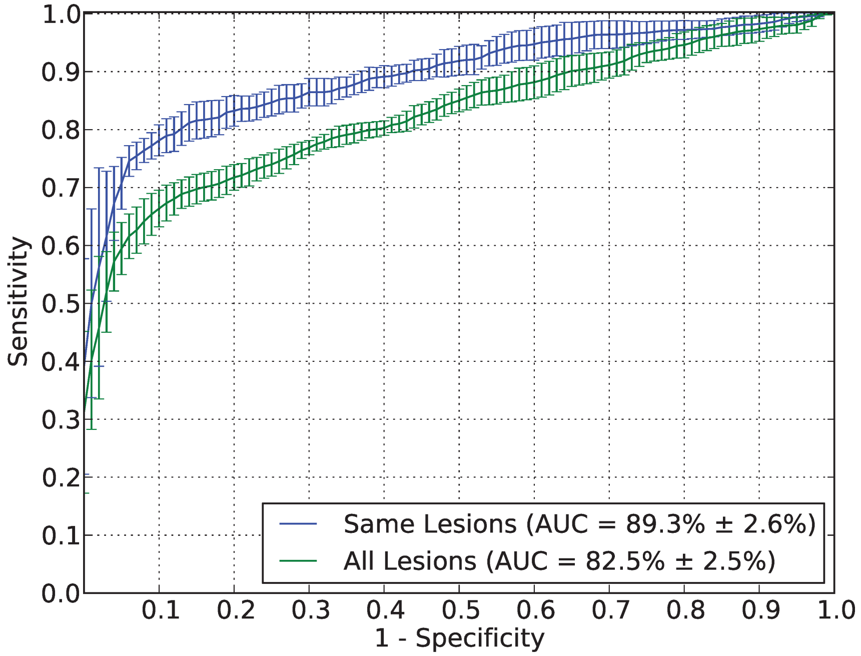


Figure 5.3: Results for the meta-classifier trained with the outcome of individual DR-based lesion detectors for DR2 dataset.

5.5 Final Remarks

A key topic of our research was that we prioritized the detection of the most common manifestation related to the disease. Given the method that provides the best results for all lesions, the sparse with semi-soft assignment as was demonstrated in the Chapter 4, we performed here a fusion technique to verify whether the patient is normal or has lesion. For this, we employed the simple method of fusion with OR and obtained promising result of 88.6% considering the detection of the most common DR-related lesions. Using the more complex fusion technique of meta-classification, which seeks a pattern based upon the scores returned by each individual DR-related lesion detector, we achieved a satisfactory AUC of 89.3%.

Chapter 6

Assessing the Need for Referral on Diabetic Retinopathy Treatment

The methods developed herein resulted in the paper submitted to a top-tier journal currently under review.

6.1 Preamble

The development of computational systems that support specialists in diverse areas of health care has been the focus of several studies [72, 40, 3, 60, 49]. The use of computational methods that aid in the diagnosis of disease have contributed significantly to improve the quality of life of patients. In this context, several computational systems have been proposed (e.g., [72, 40, 3]) for dealing complications related to a major health care problem nowadays: *Diabetes Mellitus*.

A factor that creates interest for automated screening systems is the small number of medical specialists available, in contrast to the growing number of cases of retinopathy [21].

The development of a unified screening system that identifies several different DR-related lesions simultaneously has been conducted using a bag-of-visual-words model (BoVW) based upon visual dictionaries [42, 40, 72]. However, this model needs a visual dictionaries for each type of lesion, and hence, there is a detector for each one of those types of lesion. In order to make decision such as the level of DR progression (from mild to severe), or the need for referral, one must combine those lesion detectors somehow.

In a previous study, we provided a set of five detectors for individual DR-related lesions and evaluated the performance of three different methods of fusion: logical-OR, majority vote, and meta-classification [40]. Chapter 5 also presented the results achieved for fusion with OR and meta-classification, aiming at the identification of any DR-related lesion.

As we have mentioned, due to the shrinking ratio of medical specialists/cases of retinopathy, a computational system suitable for detecting the presence of DR-related lesions in eye-fundus images is of considerable importance for the treatment. Several algorithms have been proposed for analyzing the presence/absence of retinopathy, as well as for detecting specific lesions from mild nonproliferative to proliferative retinopathy and maculopathy [38]. However, in many cases, the simple presence of a specific DR-related lesion does not represent, by itself, a reason for the patient be referred to a specialist. The presence of one or two microaneurysms, for example, may not warrant an ophthalmic specialist consultation.

In this chapter, we propose a method that can be used for assessing the need for referral, especially in remote and rural areas. The method captures retinal images from mydriatic or non-mydriatic cameras (cameras that require or not a dilatation of the pupils before the capture, respectively), evaluates the images in real-time, and suggests whether or not the patient requires a review by an ophthalmic specialist within one year after the screening. The method consists of: (1) detecting individual anomalies [72] and extracting the appropriate assessment scores (clarified in Chapter 4), and (2) classifying the image as *referable/non-referable* automatically by means of meta-classification techniques built upon the outputs of several lesion-detectors. Different from [72], we also explore alternatives for the bag-of-visual-words (BoVW) based lesion detectors, an important experimental work because the performance of BoVW depends critically on the choices of coding and pooling the low-level local descriptors aiming at characterizing the properties and signs related to each kind of lesion of interest.

The rest of the chapter is organized as follows. Section 6.2 presents the related work on image analysis. Section 6.3 explains our method of employing BoVW for creating individual detectors (individual lesion characterization), as well as the normalization techniques explored in this chapter and the meta-classification method used for combining the output of the individual detectors. Section 6.4 presents the results for the proposed approach both in terms of lesion detection as well as the referable/non-referable classification. Finally, Section 6.5 concludes the chapter.

6.2 Related work

The existence of a DR-related lesion does not necessarily indicate a vision-threatening lesion that requires a referral. The presence of microaneurysms, that characterize a moderate non-proliferative DR type, does not indicate an urgent consultation, but an indication of a follow up between three months and 12 months depending on the number and location of the microaneurysms. On the other hand, the presence of neovascularization indicates proliferative retinopathy and if not under treatment, needs urgent referral for management by an ophthalmologist [38]. Other retinal lesions that may require attention are the cotton wool spots, especially if there are more than five [87].

A nurse-managed primary care clinic is an essential step to ensure a satisfactory cost reduction as well as the opportunity of screening, assessment and treatment reaching remote communities. Nurse-led screening programs are designed to verify the presence of any DR-related lesion, as well as to identify the lesion and whether referral is required.

Screening programs for diabetic retinopathy have been developed in many countries such as the Netherlands [3], United Kingdom [60] and Australia [49]. In the Netherlands, the EyeCheck project [3] has been in operation since 2001 and more than 30,000 people with diabetes have been screened regularly between 2001 and 2010. Abràmoff et al. [3] reported a comparative study of the performance of automated DR detection using the EyeCheck, algorithm compared to the algorithm applied in the Challenge2009, winner of the 2009 Retinopathy Online Challenge Competition [70]. Evaluating the performance of the system based on retinal images of 16,670 patients, the results showed that the performance of the Challenge2009 algorithm (AUC = 82.0%) is statistically equivalent to the performance of the EyeCheck algorithm (AUC = 84.0%) [2].

The EyeCheck algorithm is based on a pixel feature classification, that is, the candidate pixels that appear to be in a red lesion. These candidate pixels are clustered in candidate lesions, from which features are extracted. These are processed with a k-NN classifier to assign a probability and to indicate the likelihood that the lesion is a red lesion [2]. The Challenge2009 algorithm uses a parametric template defined for microaneurysms. The algorithm detects the microaneurysms by locally matching a lesion template in sub-bands of wavelet transformed images, and searching for the best adapted wavelet within the lifting scheme framework [70]. Both the EyeCheck and Challenge2009 algorithms focus on the detection of specific lesions and require pre- and post-processing. In contrast to the these two algorithms, we highlight here that our method does not need any pre- and post-processing of images.

In the United Kingdon, the National Screening Committee (UK NSC) recommends a systematic population screening program to be offered annually to all people with type 1 and type 2 diabetes aged 12 or over [53]. In 2010–2011, 79% of people in England aged 12 and over identified with diabetes actually attended a retinopathy screening. In 2011–2012, this percentage increased to 81% [53, 60]. People with diabetes are invited to visit a screening venue and retinal images are captured and then graded by experts. Each image is graded for severity by a primary grader, followed by a different secondary grader. The grading outcomes for retinopathy include R0, R1 (both asked to return annually), R2 and R3 (which are referred for treatment within 13 weeks and 2 weeks, respectively) [60]. The disadvantage of the method stems from the fact that the manual grading puts a considerable burden on the health care system. Our method aims at presenting an automatic DR lesion detection [72, 40] and, assessing the necessity of a subsequent referral.

Several studies aimed at providing automated DR screening for the use by primary health care providers in rural Australian communities [49]. Luckie et al. [49] proposed the identification of proliferative retinopathy (characterized by new vessel growth). The authors exploited wave transformation, mathematical morphology operations, and fractal analysis to provide an automated assessment of images to detect vascular proliferation from one image of the macular (posterior pole) region. However, the technique requires an extensive stage of preprocessing.

Decencière et al. [22] developed a strategy to fuse a set of heterogeneous information in order to get a response about the necessity of a referral. The descriptors employed by the authors are: one pathological score per lesion (microaneurysms, exudates and hemorrhages), one signature-based pathological score (a proposed solution which relies on wavelet-based image characterizations to detect the signs of DR, and of other retinal pathologies), six quality metrics, up to nine demographic information fields and up to 18 diabetes-related information fields (age, weight, diabetes type, etc.). The heterogeneous information were fused with the algorithm for association rule mining, Apriori [5].

Our method, based on visual dictionaries is able to identify one or more different lesion types in retinal images with a unified framework. The main novelty of the current research is in the characterization of lesions using visual dictionaries and classification in referable/non-referable images, explained in the Section 6.3.

6.3 Proposed Methodology

In this section, we present the method employed to decide if a patient is to be referred to an ophthalmologist within one year after the screening. Our approach consists of (1) training detectors for individual DR-related lesions, and (2) using the scores from those detectors to train a meta-classifier that labels the retinal images as *referable* or *non-referable*. The individual detectors are based on the bags-of-visual-words (BoVW) model, for which we evaluated several possibilities of coding and pooling [12] which were explained in the Chapter 4. The meta-classification can be interpreted as the creation of a high-level feature vector of scores, for which we test three possibilities of normalization.

6.3.1 Detection of Individual DR-related Lesions

Individual DR lesions are detected using a bags-of-visual-words (BoVW) model [78, 25, 12].

As previously mentioned, the visual dictionaries are used to transform the low-level local feature vectors extracted by SURF into mid-level BoVW feature vectors. First, there is a step of *coding* in which the low-level feature vectors are given a representation in function of the dictionary. Then, for one image, all encoded vectors are aggregated in a *pooling* step (using operators such as sum, average and max).

The choice of the coding and pooling schemes has a strong impact on the performance of the BoVW representation [12]. Traditionally [78], BoVW models employed hard assignment for the coding (each local vector was assigned to its closest visual word in the dictionary, normally using Euclidean distance), and sum for the pooling. This is equivalent to create a histogram that counts the occurrences of local vectors according to their distances to the visual words. We have used this form of representation in a previous work in order to design individual lesion detectors [72].

More recently, however, both the hard-assignment and the sum-pooling have been questioned [12]. Soft assignments [47, 62, 85] have been proposed to alleviate, among other issues, the problem of boundary effects in the choice of the visual word (since a local descriptor can be more or less equidistant to several visual words). In those schemes, instead of activating the closest visual word completely, the scheme may activate several visual words, activate the visual words partially, or do both things. Similarly, the performance of pooling schemes different than the usual sum or average has been shown in several applications [12]. When soft assignment is used, the use of max pooling is especially interesting: in that case, the final mid-level vector used the maximum of each visual word activation by the local features.

In this chapter, we contrast two schemes: the common hard-sum (hard-assignment coding / sum pooling) and the more recent soft-max (soft assignment / max pooling). Both feature vectors were normalized by term-frequency (tf), which is known to result in good performances [40].

The final classification step for the individual lesion detectors is based upon a two-class *Support Vector Machine* (SVM) [11] classifier, which employs the mid-level BoVW feature vectors for training and classification. In this step, we have a binary classifier trained for each individual lesion.

6.3.2 High-level Feature Extraction

In order to decide on the referral for the patient, the information provided by each individual lesion detector is insufficient, because the lesions can be minor, they can be just a few, or they may not indicate there will be future deterioration of visual function.

Thus, our aim is at combining the evidences of the individual detectors in a meta-classification step that allows the decision-making of refer or not the patient to a doctor. This step can be interpreted as the creation of a characterization scheme based upon the classification scores of individual lesion detectors. The meta-classification is made possible by a new annotated dataset (not used in the training of any lesion detector), with images from patients tagged as referable/non-referable by an expert. That is essential for training the meta-classifier.

Here, our goal is to have a very high selectivity (very few false negatives), while also keeping high specificity (few false positives): the former is important to ensure that no patient in need stays without care, the latter is important to avoid swamping the health-care professionals with unneeded referrals.

For training the decision scheme, we proceed as follows:

1. extraction of the low-level SURF feature vectors from the training images;

2. creation of the visual dictionaries for the lesions from the annotated images of the lesion training sets;

3. extraction of the mid-level BoVW feature vectors using the visual dictionaries;

4. training of the independent lesion detectors, each with its lesion training set;

5. after training individual lesion detectors, we need to train a referable/non-referable classifier. For that, we:

   (a) extract the high-level feature vectors from the scores of the lesion detectors (step 4) on a training set of referable/non-referable tagged images;

   (b) train a meta-classifier with the high-level feature vectors from the referral/non-referral image training set.

For deciding on the referral for one particular patient, the procedure is: (1) extraction of the low-level SURF feature vectors from the retinograph images of this patient; (2) extraction of the mid-level BoVW feature vectors; (3) extraction of the high-level feature vectors from the scores of the individual SVM lesion detectors; (4) final decision based on the high-level feature vector (outcomes of the individual lesion detectors).

6.3.3 Normalization

Several problems in computer vision have benefited from fusion of several algorithms and/or sensors, with fusion in the score level being among the most used fusion approaches. Choosing the most appropriate normalization technique for the obtained scores before the fusion is a fundamentally difficult task due to the heterogeneity of the distributions of scores obtained from different data sources [75].

To verify whether the normalization methods improve on the classification outcome, we applied two simple normalization techniques: *term-frequency*, and *z-scores* (a.k.a., standard normalization). The first technique is widely used in text retrieval, where each document is represented by a vector of word frequencies [78]. Term-frequency is expressed as the division of the number of occurrences of word i in document d (n_{id}) by the total number of words in the document d (n_d). The second method, z-scores, is an adaptive score normalization based upon the Gaussian distribution. The normalized score is produced by subtracting the arithmetic mean μ of the set of scores from an original score, and dividing it by the standard deviation σ of the set of scores [75].

6.3.4 Meta-classification

As described in Section 6.3.2, individual detectors were developed for a different dataset, in order to describe the images whose aspect "necessity of referral" is known. In other words, the purpose is the development of a meta-classification system in which, in the first level, we have anomaly detectors that operate in parallel and provide an alternative (higher-level) description for each image (distances to the decision classification hyperplanes, for instance); and, in the second level, we have a classifier, trained with positive and negative images with respect to being referable or not.

Fig. 6.1(a) depicts an overview of the first level of the referable vs. non-referable approach, that consists of developing detector models for the DR-related lesions. Fig. 6.1(b) shows the referable vs. non-referable scheme which involves the creation of a high-level description, training the set of referable/non-referable tagged images, and testing step.

6.4 Validation and Experiments

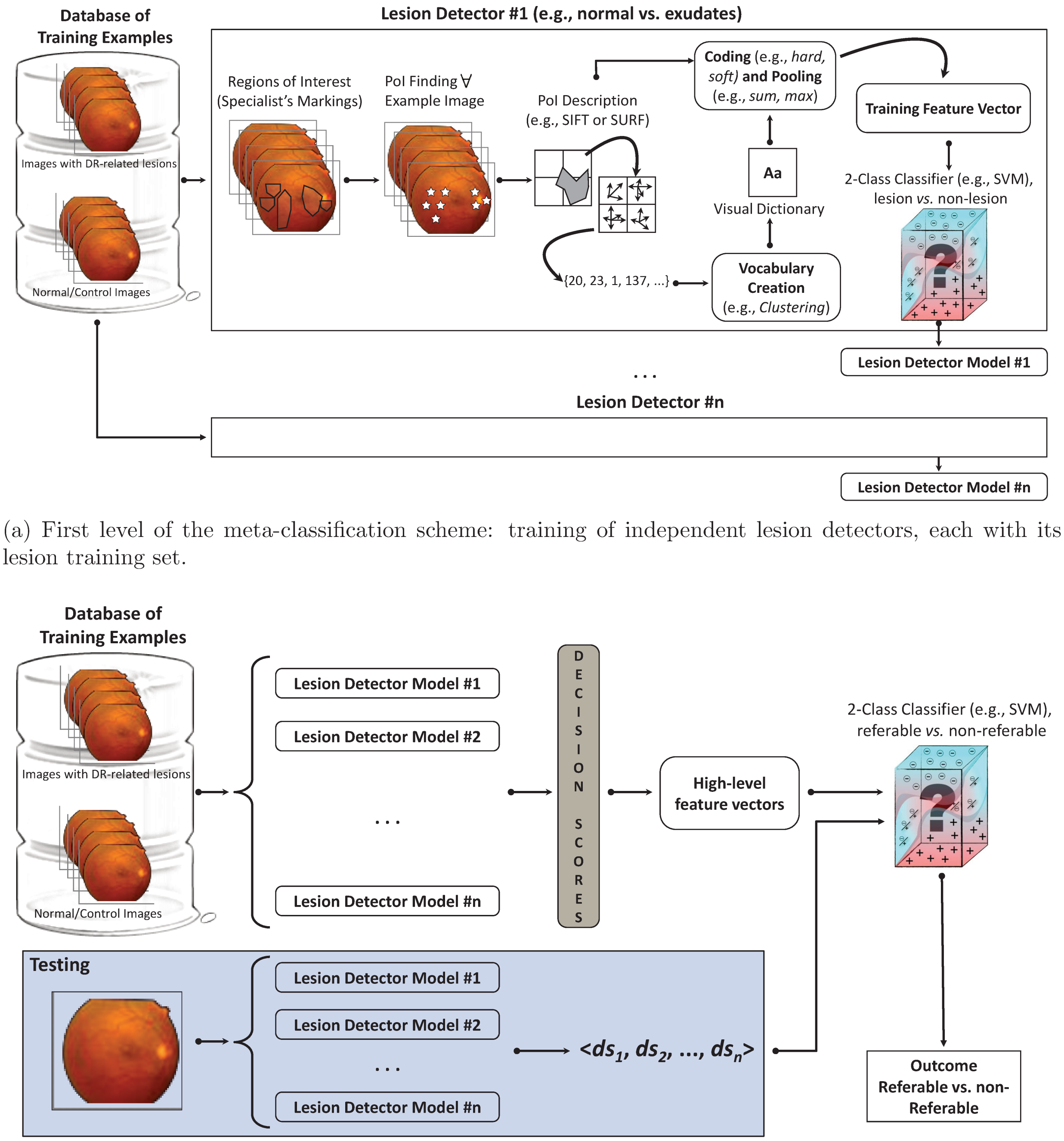
Here, we cite the datasets used in the development of the system and describe the protocol of validation employed in the meta-classification, as well as the results obtained for each experiment.

6.4.1 Datasets

We performed the experiments using two different datasets tagged by medical specialists: DR1 and DR2. Please refer to Section 4.4.1 for more information about the datasets. However, in this chapter, we are interested in the DR2 dataset labeled over other aspect: the necessity of a referral. Table 6.1 reveals the annotation occurrences for DR2 dataset.

6.4.2 5 × 2-Folds Cross-Validation

In this chapter, we have used the 5 × 2-folds cross-validation protocol [23]. The protocol consists of repeating the process of two-fold cross-validation five times. In each step,


Database of Training Examples
Images with DR-related lesions
Normal/Control Images
Lesion Detector #1 (e.g., normal vs. exudates)
Regions of Interest (Specialist's Markings)
PoI Finding ∀ Example Image
PoI Description (e.g., SIFT or SURF)
Coding (e.g., hard, soft) and Pooling (e.g., sum, max)
Training Feature Vector
Aa
Visual Dictionary
2-Class Classifier (e.g., SVM), lesion vs. non-lesion
{20, 23, 1, 137, ...}
Vocabulary Creation (e.g., Clustering)
Lesion Detector Model #1
...
Lesion Detector #n
Lesion Detector Model #n
Database of Training Examples
Images with DR-related lesions
Normal/Control Images
Lesion Detector Model #1
Lesion Detector Model #2
...
Lesion Detector Model #n
DECISION SCORES
High-level feature vectors
2-Class Classifier (e.g., SVM), referable vs. non-referable
Testing
Lesion Detector Model #1
Lesion Detector Model #2
...
Lesion Detector Model #n
<ds1, ds2, ..., dsn>
Outcome Referable vs. non-Referable

(a) First level of the meta-classification scheme: training of independent lesion detectors, each with its lesion training set.

(b) Second level of the meta-classification scheme: extracting the high-level feature vectors from the scores of the lesion detectors (using the models described in Fig. 6.1(a)) on a training set of referable/non-referable tagged images. Training a meta-classifier with the high-level feature vectors from the referral/non-referral image training set.

Figure 6.1: An overview of the referable vs. non-referable approach.

Table 6.1: Annotation occurrences for DR2 dataset

Need be referred	DR2
Positive	98
Negative	337

the dataset is randomly divided in two groups. The first group is used as training set and the second group as test set. Then, the groups are switched. We use the 5 × 2-folds cross-validation because it shows to be slightly more powerful than other validation protocols [23].

6.4.3 Experiments

Here, we present the results for evaluating retinal images whether the patient needs to be referred to an ophthalmologist or not within one year. There are several metrics to measure the performance of an algorithm for detection/classification.

The experiments are divided in two parts:

1. **Part #1.** Experiments for detecting individual anomalies. Here, we use a cross-dataset validation, training the classifiers with DR1 and testing with DR2.

2. **Part #2.** Experiments for determining the necessity of referral using the scores obtained for the images from the DR2 dataset and the lesion classifiers trained on DR1 dataset.

Experiments - Part #1

The initial identification of lesions was conducted in previous research and published in [72, 40]. In previous work, we explored only the hard-sum as mid-level feature extraction. Here our objective is to create a framework that is able to assign a score to retinal images and testing the detectors with the DR2 dataset. Therefore, we also explore alternatives to the mid-level feature extraction such as the soft-max.

For the binary classification technique, we used the SVM classifier. We searched for the best SVM parameters during training using the standard LibSVM's grid search fine tuning algorithm [16].

Table 6.2 shows the results obtained by the individual detectors. The results indicate that the soft-max technique has a considerable advantage compared to hard-sum for detecting white lesions (hard exudates and cotton wool spots), except for drusen. Hard-sum performed better for red lesion detection (deep hemorrhages and the complete set of red lesions). This complementary results goes in line with recent studies in the Computer

Vision literature which hints at the interesting properties and results of soft-max techniques [12, 47, 62, 85]. Note that the results correspond to those presented in the Table 4.3 for hard and soft, and for the respective lesions.

Table 6.2: AUCs for individual detectors

DR-related Lesion	**Hard-sum**	**Soft-max**
Hard Exudates	93.1%	**95.5%**
Superficial Hemorrhages	88.8%	88.7%
Deep Hemorrhages	**90.0%**	86.5%
Red Lesions	**92.3%**	87.1%
Cotton Wool Spots	82.1%	**84.9%**
Drusen	**66.5%**	62.6%

Experiments - Part #2

In the second part of the experiments, we evaluated the referable/non-referable meta-classifier we propose. For that, we extracted the scores for the classification of the retinal images, generating a high-level description. This process is performed through the fusion by meta-classification, where the outputs of a series of individual lesion classifiers are used as input to a new classifier at a higher level. We explore how this meta-classifier performs when fed with normalized and non-normalized classification scores from the first level.

- *Without Normalization:* The experiments without normalization explore the raw scores generated by the individual lesion detectors to distinguish between referable and non-referable images. As explained in 6.3.1, we describe each image with the visual word dictionaries previously constructed for each DR-related lesion, feed the respective lesion detector with the created feature vector, and obtain the score. This procedure is repeated for each lesion, creating a final feature vector of six dimensions (each dimension refers to the output of an individual lesion detector). Then, we run the meta-classification with the discussed cross-validation protocol. Figs. 6.2 and 6.3 depict the results for hard-sum and soft-max, respectively. Note that, in general, the soft-max yielded the best results.

- *Normalization with term-frequency:* For comparison purposes, after the creation of the feature vectors based on the classification scores of individual lesion detectors, we applied two normalization techniques. The first technique is the term-frequency, common in text retrieval. With the final feature vectors constructed, we proceed with the classification as explained earlier. Figs. 6.4 and 6.5 show the ROC classification results for hard-sum and soft-max, respectively.

- *Normalization with z-score:* As a last normalization technique, we employed the z-score technique in this work. Given the six scores of a specific retinal image, to obtain the final feature vector, we begin calculating the arithmetic mean μ and the standard

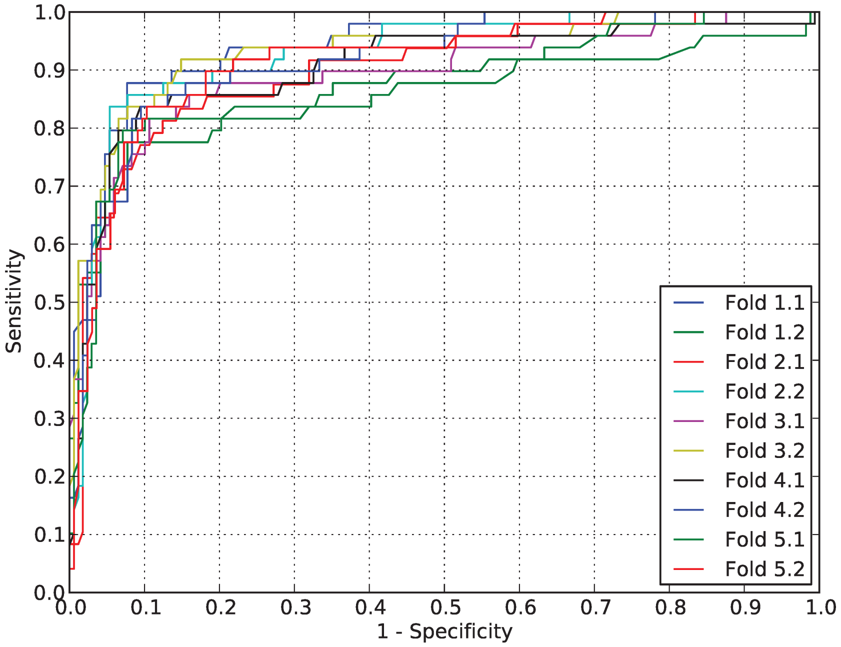


Figure 6.2: ROC assessing the selectivity and sensitivity for the final referral decision, when using hard-sum mid-level BoVW feature vectors, and no normalization for the high-level feature vector of scores. Cross-validation with 5 × 2-folds.

deviation σ of the scores of a feature vector. Then, we subtract μ of each score and divide it by σ. Then, we perform the classification. Figs. 6.6 and 6.7 illustrate the AUCs obtained for hard-sum and soft-max, respectively considering z-score normalization.

Table 6.3 summarizes all the results obtained for referral, presenting the arithmetic mean and standard deviation. In an analysis of the results of the hard-sum technique, we can note that using the normalization with term-frequency (AUC = 82.5%) is a drawback. The result fell more than eight percentage points in comparison with the method without normalization. On the other hand, the z-score technique reached an area under the curve equal to 91.7%, higher than the result without normalization. However, the standard deviations made the methods statistically equivalent, therefore normalization does not improve on the classification outcome and does not have to be considered for the problem and data we discuss in this chapter.

Fig. 6.8 depicts the curves which present the mean and standard deviation for the hard-sum approach. In this case, note that z-score is also equivalent to not using normalization.

As for the soft-max approach, the disadvantage of the term-frequency technique is maintained. The result without normalization, 93.4%±2.1%, is significantly higher than

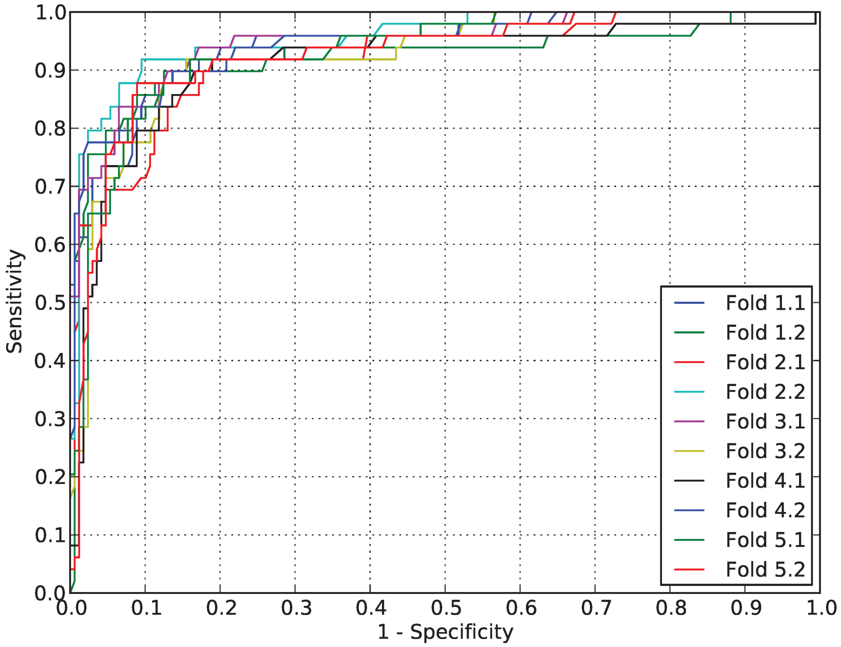


Figure 6.3: ROC assessing the selectivity and sensitivity for the final referral decision, when using soft-max mid-level BoVW feature vectors, and no normalization for the high-level feature vector of scores. Cross-validation with 5 × 2-folds.

Table 6.3: AUCs for referral

Technique	**Hard-sum**	**Soft-max**
Without normalization	90.8%±3.1%	**93.4%±2.1%**
Term-frequency	82.5%±4.6%	83.4%±4.6%
Z-score	**91.7%±2.1%**	89.4%±3.0%

the result with z-score normalization technique, 89.4%±3.0%. The ROC curves for arithmetic mean and standard deviation from the soft-max approach are depicted in Fig. 6.9.

Contrasting the best results for hard-sum and soft-max, highlighted in Table 6.3, we note a difference of 1.7 percentage points for the arithmetic mean, and the same standard deviation. This does not characterize a considerable advantage for the method which does not explore any normalization technique, but shows that the normalization is not worth to increase the final classification accuracy. Furthermore, although the normalization with z-score provides an approximate outcome, the extra computation burden for computing the mean and standard deviation makes the method more expensive than its counterpart with no normalization.

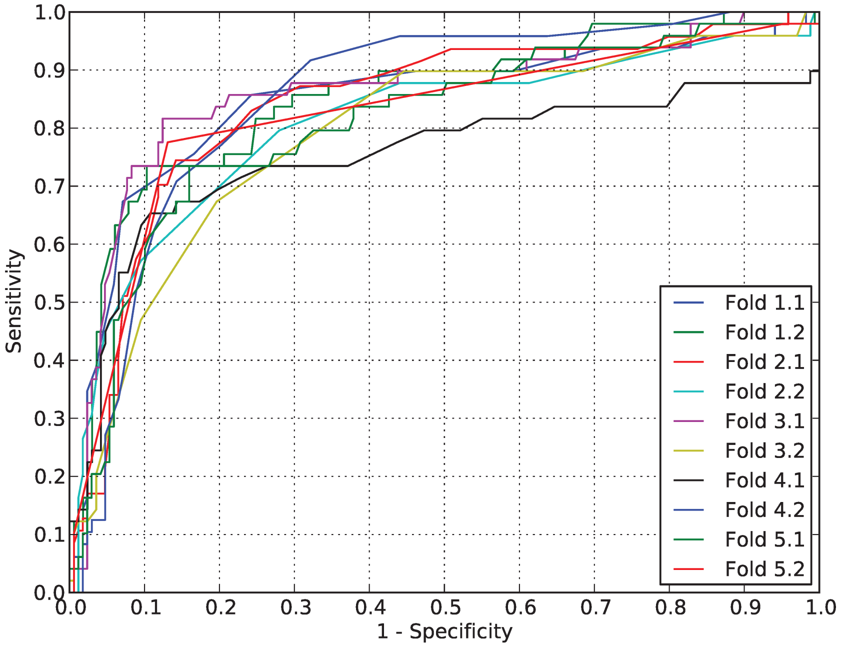


Figure 6.4: ROC assessing the selectivity and sensitivity for the final referral decision, when using hard-sum mid-level BoVW feature vectors, and term-frequency normalization for the high-level feature vector of scores. Cross-validation with 5 × 2-folds.

Comparison with the State of the Art

In a strategy similar to ours, Decencière et al. [22] combined scores extracted by individual DR-related lesion detectors in order to identify whether the patient needs or not to be referred to a specialist. However, it is important to emphasize some crucial differences in our work: (1) the authors use parameters of quality assessment as descriptors, while we perform the analysis in a previous step (see Chapters 2 and 3); (2) we do not explore contextual information; and (3) we perform a cross-dataset training/testing, a setup closer to a real operation scenario. Fig. 6.10 depicts the ROC curve obtained by the authors for the classification as "normal"/"to be referred to a specialist". In comparison to our results with soft-max approach (Fig. 6.9), we can observe a high advantage of our method. For example, for a sensitivity of 90.0%, the authors have a specificity of 50.0%, while we have 85. 0%. In other words, Decencière et al. save the specialist time (avoiding to attend normal people) in 50.0%, while we save in 85%.

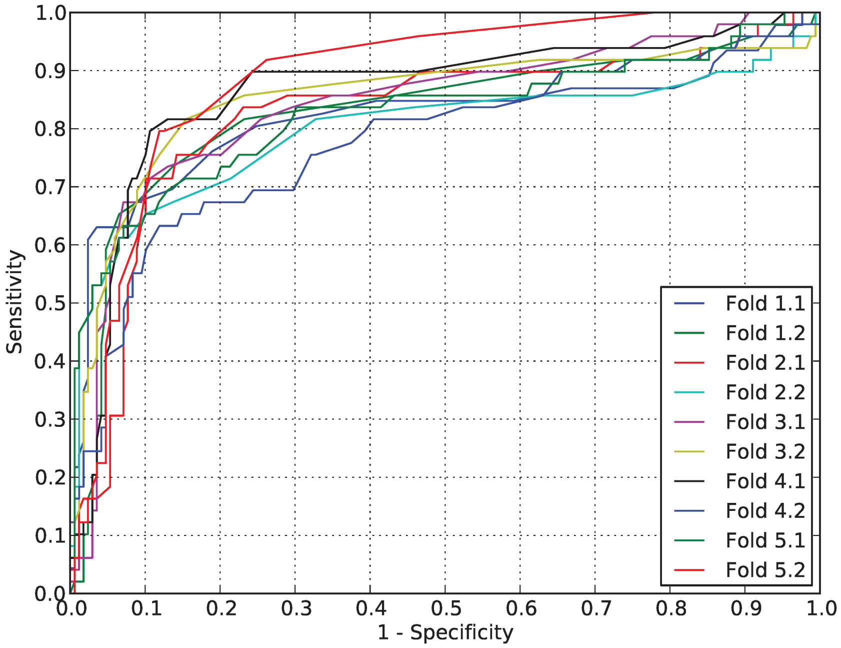


Figure 6.5: ROC assessing the selectivity and sensitivity for the final referral decision, when using soft-max mid-level BoVW feature vectors, and term-frequency normalization for the high-level feature vector of scores. Cross-validation with 5 × 2-folds.

6.5 Final Remarks

The increasing prevalence of diabetes, and associated diabetic retinopathy, contrasted to a shortage of available specialists, poses a challenge, especially in rural and remote areas. That challenge has stimulated the research and development of smart computational systems to increase the availability of screening, without wasting resources in unnecessary referrals to ophthalmological services. Published outcomes of those efforts include the detection of interest regions of the retina such as the (e.g., optic disc and macula), the assessment of image quality, and the analysis of presence/absence of DR-related lesions and other pathologies.

In a previous work, we used Machine Learning and Computer Vision techniques to detect DR-related lesions such as hard exudates, superficial hemorrhages, deep hemorrhages, cotton wool spots, drusen, and red lesions [40]. However, previously we only investigated the most traditional BoVW coding/pooling technique, hard-assignment for coding, and sum for pooling (hard-sum). In this chapter, we explore other alternatives, which is important because recent literature on the BoVW model has shown that the

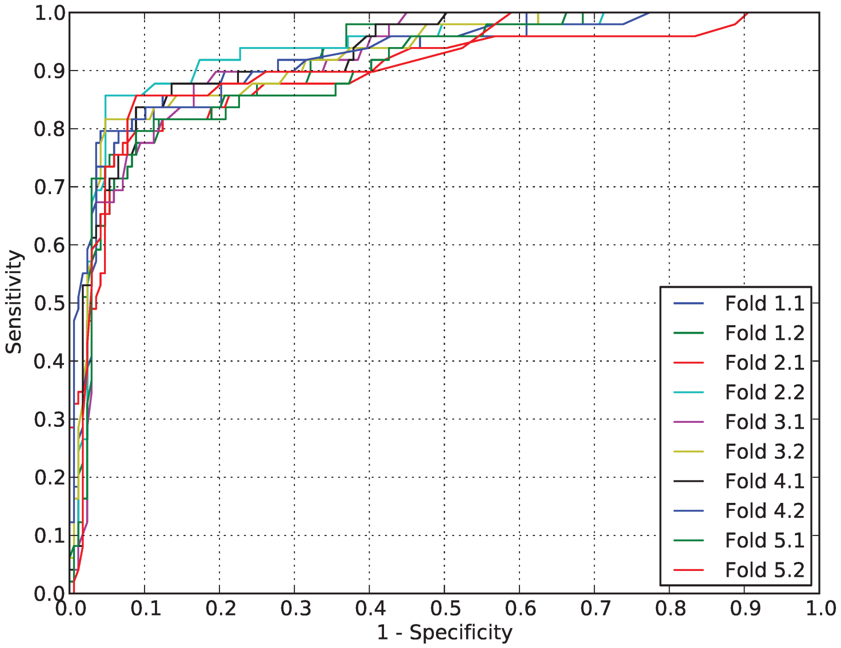


Figure 6.6: ROC assessing the selectivity and sensitivity for the final referral decision, when using hard-sum mid-level BoVW feature vectors, and z-scores (standard normalization) for the high-level feature vector of scores. Cross-validation with 5 × 2-folds.

coding/pooling choice has much impact in its performance. Our experiments show that, although the more recent soft-max scheme performs better for some of the lesions, the traditional hard-sum scheme is better for others, indicating that aided-diagnostic datasets have specificities unlike the ones of generic image recognition, like the ones evaluated by Boureau et al. [12].

The main novelty of this chapter is the use of fusion by meta-classification as a powerful tool for deciding whether a retinal image warrants a referral for the patient attending an appointment with a doctor within one year after screening. This meta-classification approach is based on the development of a new meta-classifier trained and tested with the scores generated by the individual lesion detectors. The scheme may be interpreted as the extraction of a high-level feature, composed of the scores of the individual lesion detectors, which is then used in the training and the decision in the meta-classifier.

We have also explored normalization techniques on the high-level feature vector of scores, comparing the performance of the raw scores, term-frequency normalized scores, and z-scores (standard normalized scores). Contrarily to our expectations, normalization did not lead to better results: term-frequency actually reduced the performance, while

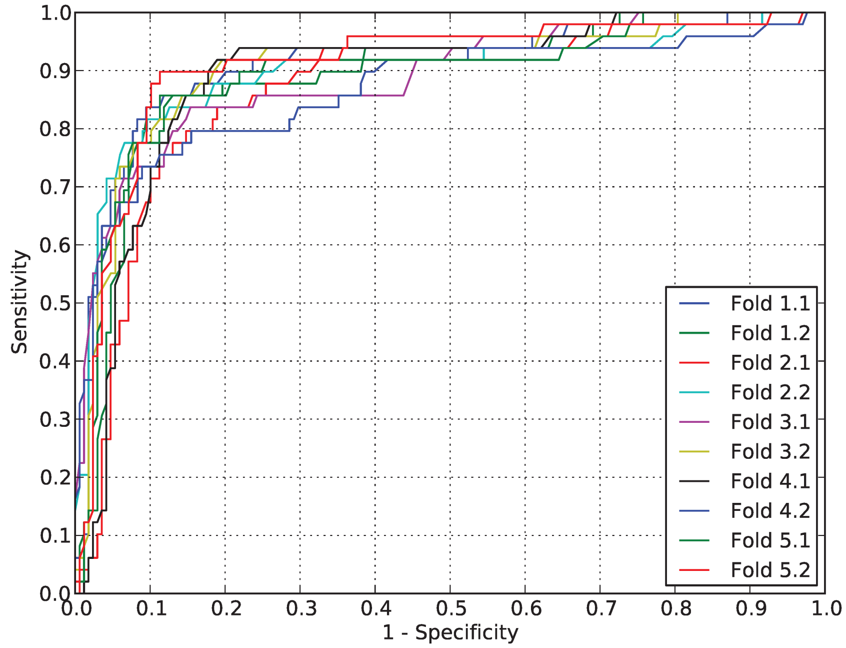


Figure 6.7: ROC assessing the selectivity and sensitivity for the final referral decision, when using soft-max mid-level BoVW feature vectors, and z-scores (standard normalization) for the high-level feature vector of scores. Cross-validation with 5 × 2-folds.

the z-scores were statistically tied to the raw scores but imposing extra computation operations. A possible explanation to that phenomenon was the sensitivity of normalization techniques to noise.

The best result achieved by our approach reached an area under the ROC curve (AUC) of 93.4%, for the high-level feature vector of scores. Those results show the potential for a big impact, since they indicate that our technique can already be used quite successfully to indicate the need for referral specially when we consider that we are using a unified framework for lesion detection and using a cross-training/testing protocol for the developed methodologies. For the unified framework, we mean that any new detector could be simply connected to the framework by creating its appropriate visual dictionary without the need for any pre- or post-processing operation. For the cross-training/testing protocol, we mean that using our method it is possible to train the individual lesion detectors on one kind of image and test in images with different acquisition conditions (e.g., changing device, operator, etc.).

While operating alone, the lesion detectors could be providing unnecessary doctor appointments. In this sense, the meta-classifier is able to combine the evidences in an

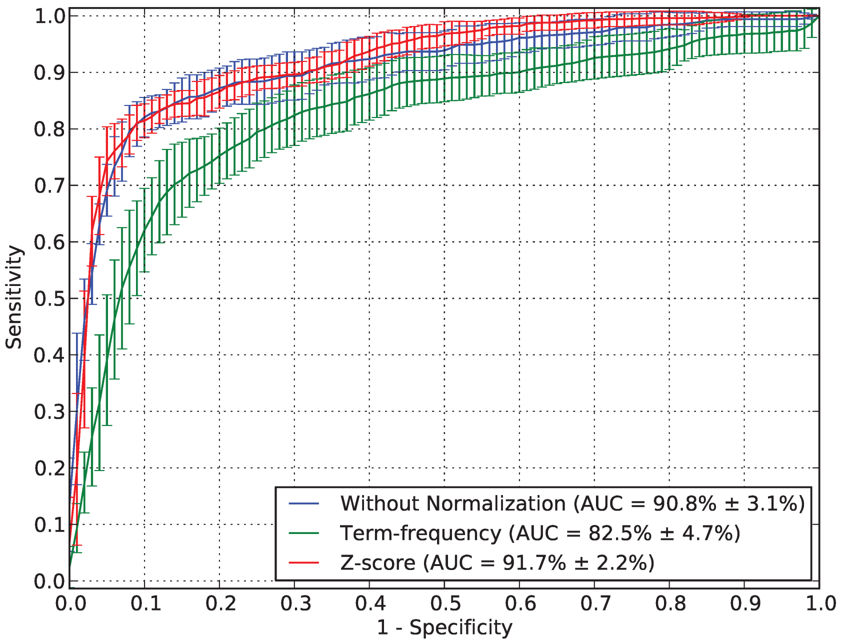


Figure 6.8: Arithmetic mean and standard deviation of the ROC curves for the final referral decision, when using hard-sum mid-level BoVW features considering term-frequency and z-score normalization techniques as well as no-normalization.

effective way optimizing available resources.

Finally, we were surprised by the results of the normalization, thus, an interesting future work would be testing robust (insensitive to noise) normalization techniques, such as the w-score proposed by Scheirer et al. [75].

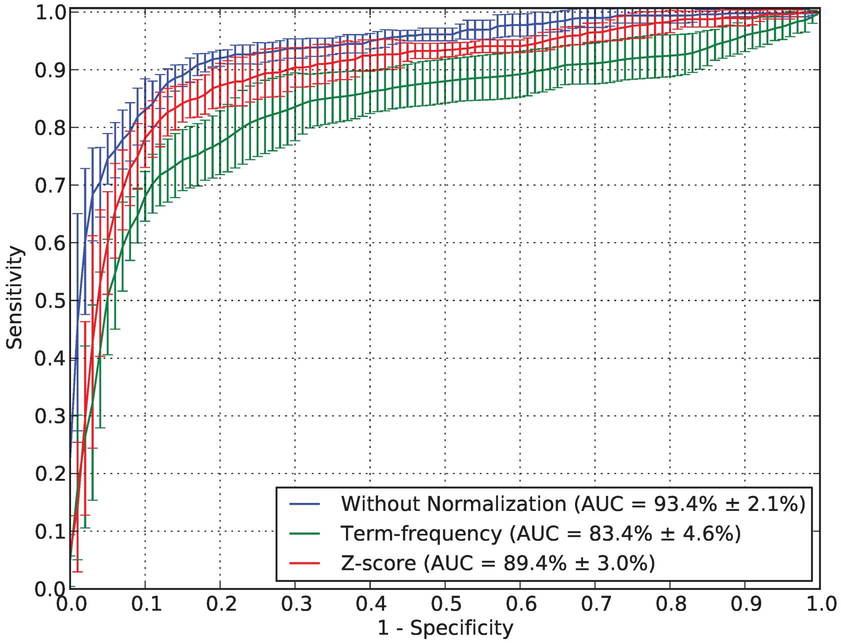


Figure 6.9: Arithmetic mean and standard deviation of the ROC curves for the final referral decision, when using soft-max mid-level BoVW features considering term-frequency and z-score normalization techniques as well as no-normalization.

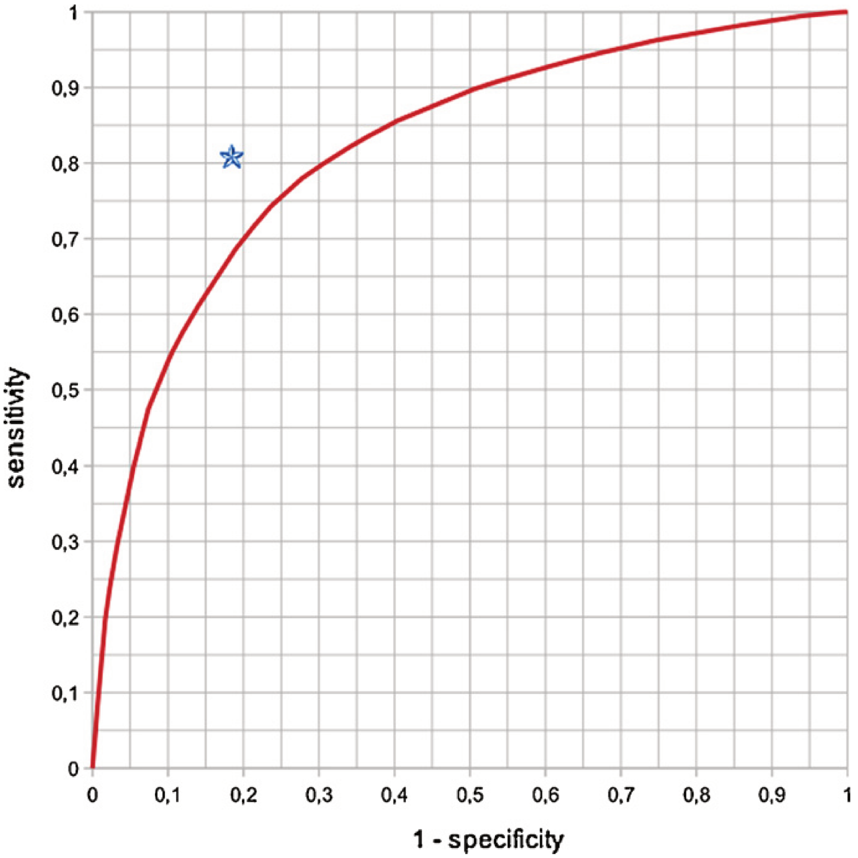


Figure 6.10: ROC curve resulting from [22], obtained on a specific dataset (different from ours). The blue star shows the result obtained by a human expert on a subset extracted randomly from the same dataset. Note that this human expert was not exactly in the same conditions of the interpretation of the system.

Chapter 7

Conclusions

In this work, presented as a mixed format of papers' collection, we proposed a general framework to automate the retinal image analysis. Our contribution is a set of methods which ranges from the crucial step immediately after the capture of an image until the verification of presence/absence of any diabetic retinopathy related lesion, as well as the evaluation of the necessity of referral of a patient to a medical specialist.

Our work yields a considerable contribution to the Computer Vision literature with the proposal of sparse feature extraction + semi-soft coding combination, which works particularly well for the DR-related lesion detection challenge.

Sufficient quality is a necessary prerequisite on input images for reliable automatic detection systems in several healthcare environments. **Image quality assessment** represents an important limiting factor for automated DR screening. The assessment of retinal image quality is a critical step to obtain satisfactory and reliable outcomes in a screening system. **Chapters 2** and **3** discussed a set of proposed approaches, exploring the existing techniques in the literature for image quality evaluation and presenting innovations in this context.

In the main work performed for the quality analysis, we discussed image quality regarding two aspects: *field definition* and *blur analysis*. For field definition, we proposed the use of structural similarity measures to evaluate the quality of retinal images. The method deployed for blur detection involves a series of different blurring classifiers, relying upon four descriptors: vessel area, visual dictionaries, progressive blurring and progressive sharpening. The main breakthrough for this step was the use of classifier fusion to optimize the classification. This tactic gave us an interesting result: to ensure that a satisfactory percentage of poor quality images will be discovered ($\simeq$ 90.0%), we can establish that only 10.0% of the enough quality images will be unnecessarily retaken. The quality assessment constitutes a key step of a robust DR-related lesion screening system because it helps preventing misdiagnosis and posterior retake.

Detection of individual DR-related lesions is one of the most important topics of this work. Many works have focused on the detection of individual lesions, aiming at facilitating the attendance in rural and remote communities. The individual DR-related lesion detectors were implemented as a projection to the development of a more complex approach to represent retinal images, in a high-level assignment stage. This strategy provided us a robust representation method for the eye-fundus images and opened the opportunity to contribute deeply to the Medical literature, for both multi-lesion detection and for referral.

The results achieved for individual DR-related lesion detection, as well as a minutely explanation of the bag of visual words representation (explored in all steps of our general work), are contained in **Chapter 4**. In that chapter, we explored several alternatives for the extraction of low-level features, and the creation of mid-level representations pointing out important choices when designing a unified framework for detecting DR lesions. The high-level features, characterized by the scores extracted of each DR-related lesion detection, constitutes our fundamental tool for the next steps.

A considerable contribution of this step was the proposal of a new coding scheme called *semi-soft*, which explores the advantages of the most traditional hard-sum and soft-max coding/pooling assignments. With a detailed experimental design evaluation through analysis of variance (ANOVA), we showed that the semi-soft coding associated with sparse feature extraction provided a good balance for designing an efficient and effective DR-related lesion detector with results that outperform the ones in the literature.

Based upon the scores associated to the detection of the most common DR-related lesion, we developed an accurate **multi-lesion detector** which showed to be effective for the detection of all the considered lesions. The strategy deployed for the multi-lesion detector construction was the fusion of individual lesion detectors. We used two techniques, logical OR and meta-classification. This latter provided us the most satisfactory result. A description of the multi-lesion detector was shown in **Chapter 5**.

Finally, aimed at providing a tool which allows to save the time of the ophthalmologists and to guarantee that patients who need urgent referral have priority, we took advantage of the same classification scores extracted from the individual DR-related lesion detectors and use them as high-level features for evaluating whether or not a patient needs to be referred to an ophthalmologist. Some works have been developed with this objective exploring distinct methods, and our method showed considerable advantages over them.

Our method, described in **Chapter 6**, can be used for assessing the need for **referral**, especially in remote and rural areas. The method captures retinal images from non-mydriatic or mydriatic cameras, evaluates the images in real-time, and suggests whether or not the patient requires a review by an ophthalmic specialist within one year after the screening. We have achieved important results with the proposed methodology. For example, for a sensitivity of 90.0%, we have a specificity of 85.0%, which means that the specialist time may be saved in 85.0% (only 15.0% of the attended patients will be normal).

In closing this work, we would like to emphasize that there is still important researches to be done in DR image analysis. For instance, identifying the precise location of the lesion, as well as the size and quantity, and defining the DR severity degree of a patient further classifying the images as related to DR cases in early, mild, proliferative and severe stages.

Bibliography

[1] Michael D. Abràmoff, Meindert Niemeijer, Maria S. A. Suttorp-Schulten, Max A. Viergever, Stephen R. Russell, and Bram van Ginneken. Evaluation of a system for automatic detection of diabetic retinopathy from color fundus photographs in a large population of patients with diabetes. *Diabetes Care*, 31(2):193–198, 2008.

[2] Michael D. Abràmoff, Joseph M. Reinhardt, Stephen R. Russell, James C. Folk, Vinit B. Mahajan, Meindert Niemeijer, and Gwénolé Quellec. Automated early detection of diabetic retinopathy. *Ophthalmology*, 117(6):1147–1154, 2010.

[3] Michael D. Abràmoff and Maria S. A. Suttorp-Schulten. Web-based screening for diabetic retinopathy in a primary care population: the eyecheck project. *Telemedicine Journal & e-Health*, 11:668–674, 2005.

[4] Udyavara R. Acharya, Choo M. Lim, E. Yin Kwee Ng, Caroline Chee, and Toshiyo Tamura. Computer-based detection of diabetes retinopathy stages using digital fundus images. *Journal of Engineering in Medicine*, 223:545–553, 2009.

[5] Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules in large databases. In *Intl. Conference on Very Large Data Bases*, pages 487–499, 1994.

[6] Carla Agurto, Victor Murray, Eduardo Barriga, Sergio Murillo, Marios Pattichis, Herbert Davis, Stephen Russell, Michael Abràmoff, and Peter Soliz. Multiscale am-fm methods for diabetic retinopathy lesion detection. *IEEE Transactions on Medical Imaging*, 29(2):502–512, 2010.

[7] Sandra Avila, Nicolas Thome, Matthieu Cord, Eduardo Valle, and Arnaldo de A. Araújo. Pooling in image representation: The visual codeword point of view. *Computer Vision and Image Understanding*, 117(5):453–465, 2013.

[8] Ricardo Baeza-Yates and Berthier Ribeiro Neto. *Modern Information Retrieval*, volume 1. Addison Wesley, 1999.

[9] Adam Baumberg. Reliable feature matching across widely separated views. In *IEEE Intl. Conference on Computer Vision and Pattern Recognition*, volume 1, pages 774–781. IEEE, 2000.

[10] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (SURF). *Computer Vision and Image Understanding*, 110(3):346–359, 2008.

[11] Christopher M. Bishop. *Pattern Recognition and Machine Learning.* Springer, 1 edition, 2006.

[12] Y-Lan Boureau, Francis Bach, Yann LeCun, and Jean Ponce. Learning mid-level features for recognition. In *IEEE Intl. Conference on Computer Vision and Pattern Recognition*, pages 2559–2566, 2010.

[13] Y-Lan Boureau, Jean Ponce, and Yann LeCun. A theoretical analysis of feature pooling in visual recognition. In *Intl. Conference on Machine Learning*, pages 111–118, 2010.

[14] Peter Bragge, Russell L. Gruen, Marisa Chau, Andrew Forbes, and Hugh R. Taylor. Screening for Presence or Absence of Diabetic Retinopathy: A Meta-analysis. *Archives of Ophthalmology*, 129(4):435–444, 2011.

[15] Phillip K. Chan and Salvatore J. Stolfo. Toward parallel and distributed learning by meta-learning. In *AAAI Workshop in Knowledge Discovery in Databases*, pages 227–240, 1993.

[16] Chih-Chung Chang and Chih-Jen Lin. *LIBSVM: a library for support vector machines*, 2001. Software at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[17] David J. Cornforth, Herbert F. Jelinek, Malvin C. Teich, Steven B. Lowen, et al. Wrapper subset evaluation facilitates the automated detection of diabetes from heart rate variability measures. In *Intl. Conference on Computational Intelligence for Modelling Control and Automation*, pages 446–455, 2004.

[18] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[19] Michael J. Cree, E. Gamble, and David J. Cornforth. Colour normalisation to reduce inter-patient and intra-patient variability in microaneurysm detection in colour retinal images. In *Workshop on Digital Image Computing*, pages 163–168, 2005.

[20] Herbert Davis, Stephen Russell, Eduardo Barriga, Michael D. Abramoff, and Peter Soliz. Vision-based, real-time retinal image quality assessment. In *IEEE Intl. Computer-Based Medical Systems*, pages 1–6, 2009.

[21] Tiago José de Carvalho. Aplicação de técnicas de visão computacional e aprendizado de máquina para a detecção de exsudatos duros em imagens de fundo de olho. Master's thesis, University of Campinas, 2010.

[22] Étienne Decencière, Guy Cazuguel, X Zhang, Guillaume Thibault, J-C Klein, Fernand Meyer, Beatriz Marcotegui, Gwénolé Quellec, M Lamard, R. Danno, et al. Teleophta: Machine learning and image processing methods for teleophthalmology. *Ingénierie et Recherche Biomédicale*, 34(2):196–203, 2013.

[23] Thomas G. Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10:1895–1923, 1998.

[24] Ioannis N. Dimou, Georgios C. Manikis, and Michalis E. Zervakis. Classifier fusion approaches for diagnostic cancer models. In *IEEE Engineering in Medicine and Biology Society*, pages 5334–5337. IEEE, 2006.

[25] Eduardo Alves do Valle Jr. *Local-Descriptor Matching for Image Identification Systems*. PhD thesis, Université de Cergy-Pontoise École Doctorale Sciences et Ingénierie, Cergy-Pontoise, France, June 2008.

[26] Karen M. Facey, National Health Service in Scotland, and Health Technology Board for Scotland. *Health Technology Assessment: Organisation of services for diabetic retinopathy screening*. Health Technology Board for Scotland, 2002.

[27] Pedro F. Felzenszwalb, Ross B. Girshick, and David McAllester. Cascade object detection with deformable part models. In *IEEE Intl. Conference on Computer vision and Pattern Recognition*, pages 2241–2248. IEEE, 2010.

[28] Alan D. Fleming, Sam Philip, Kate A. Goatman, John A. Olson, and Peter F. Sharp. Automated assessment of diabetic retinal image quality based on clarity and field definition. *Investigative Ophthalmology & Visual Science*, 47(3):1120–1125, 2006.

[29] Alan D. Fleming, Sam Philip, Kate A. Goatman, John A. Olson, and Peter F. Sharp. Automated microaneurysm detection using local contrast normalization and local vessel detection. *IEEE Transactions Medical Imaging*, 25:1223–1232, 2006.

[30] Alan D. Fleming, Sam Philip, Keith A. Goatman, Graeme J. Williams, John A. Olson, and Peter F. Sharp. Automated detection of exudates for diabetic retinopathy screening. *Physics in Medicine and Biology*, 52(24):7385–7396, 2007.

[31] Luc M. J. Florack, Bart M. Ter Haar Romeny, Jan J. Koenderink, and Max A. Viergever. General intensity transformations and differential invariants. *Journal of Mathematical Imaging and Vision*, 4(2):171–187, 1994.

[32] G. G. Gardner, David L. Keating, Tom H. Williamson, and Alex T. Elliott. Automatic detection of diabetic retinopathy using an artificial neural network: a screening tool. *British Journal of Ophthalmology*, 80(11):940–944, 1996.

[33] Luca Giancardo, Fabrice Meriaudeau, Thomas P. Karnowski, Edward Chaum, and Kenneth Tobin. *New Developments in Biomedical Engineering*, chapter Quality Assessment of Retinal Fundus Images using Elliptical Local Vessel Density, pages 201–224. InTech, 2010.

[34] Luca Giancardo, Fabrice Meriaudeau, Thomas P. Karnowski, Yaqin Li, Kenneth Tobin, and Edward Chaum. Microaneurysm detection with radon transform-based classification on retina images. In *Intl. Conference of the IEEE Engineering in Medicine and Biology Society*, pages 5939–5942, 2011.

[35] Michael H. Goldbaum, Pamela A. Sample, Kwokleung Chan, Julia Williams, Te-Won Lee, Eytan Blumenthal, Christopher A. Girkin, Linda M. Zangwill, Christopher Bowd, Terrence Sejnowski, et al. Comparing machine learning classifiers for diagnosing glaucoma from standard automated perimetry. *Investigative Ophthalmology & Visual Science*, 43(1):162–169, 2002.

[36] Rafael C. Gonzalez and Richard E. Woods. *Digital Image Processing.* Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 2nd edition, 2006.

[37] Anil K. Jain, Robert P. W. Duin, and Jianchang Mao. Statistical pattern recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):4–37, 2000.

[38] Herbert F. Jelinek and Michael J. Cree. *Automated Image Detection of Retinal Pathology.* Taylor & Francis, 2010.

[39] Herbert F. Jelinek, Michael J. Cree, David Worsley, Alan P. Luckie, and Peter Nixon. An automated microaneurysm detector as a tool for identification of diabetic retinopathy in rural optometric practice. *Clinical and Experimental Optometry*, 89(5):299–305, 2006.

[40] Herbert F. Jelinek, Ramon Pires, Rafael Padilha, Siome Goldenstein, Jacques Wainer, and Anderson Rocha. Data fusion for multi-lesion diabetic retinopathy detection. In *IEEE Intl. Computer-Based Medical Systems*, pages 1–4, 2012.

[41] Herbert F. Jelinek, Ramon Pires, Rafael Padilha, Siome Goldenstein, Jacques Wainer, and Anderson Rocha. Quality control and multi-lesion detection in automated retinopathy classification using a visual words dictionary. In *Intl. Conference of the IEEE Engineering in Medicine and Biology Society*, 2013.

[42] Herbert F. Jelinek, Anderson Rocha, Tiago Carvalho, Siome Goldenstein, and Jacques Wainer. Machine learning and pattern classification in identification of indigenous retinal pathology. In *Intl. Conference of the IEEE Engineering in Medicine and Biology Society*, pages 5951–5954, 2011.

[43] Josef Kittler, Mohamad Hatef, Robert P. W. Duin, and Jiri Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239, 1998.

[44] Ludmila I. Kuncheva. Combining pattern classifiers: Methods and algorithms. *IEEE Transactions on Neural Networks*, 18(3):964–964, 2007.

[45] Marc Lalonde, Langis Gagnon, and Marie-Carole Boucher. Automatic visual quality assessment in optical fundus images. *Proceedings of Vision Interface*, pages 259–264, 2001.

[46] Yaqin Li, Thomas P. Karnowski, Kenneth W. Tobin, Luca Giancardo, Scott Morris, Sylvia E. Sparrow, Seema Garg, Karen Fox, and Edward Chaum. A health insurance portability and accountability act–compliant ocular telehealth network for the remote diagnosis and management of diabetic retinopathy. *Telemedicine and e-Health*, 17(8):627–634, 2011.

[47] Lingqiao Liu, Lei Wang, and Xinwang Liu. In defense of soft-assignment coding. In *IEEE Intl. Conference on Computer Vision*, pages 2486–2493, 2011.

[48] David G. Lowe. Distinctive image features from scale-invariant keypoints. *Intl. Journal of Computer Vision*, 60(2):91–110, 2004.

[49] Alan P. Luckie, Herbert F. Jelinek, Michael J. Cree, R. Cesar, J. Leandro, C. McQuellin, P. Mitchell, et al. Identification and follow-up of diabetic retinopathy in rural health in australia: an automated screening model. *Investigative Ophthalmology & Visual Science*, 45(5):5245, 2004.

[50] David Maberley, Andrew Morris, Dawn Hay, Angela Chang, Laura Hall, and Naresh Mandava. A comparison of digital retinal image quality among photographers with different levels of training using a non-mydriatic fundus camera. *Ophthalmic Epidemiology*, 11(3):191–197, 2004.

[51] Quresh Mohamed, Mark C. Gillies, and Tien Y. Wong. Management of diabetic retinopathy. *JAMA: the Journal of the American Medical Association*, 298(8):902–916, 2007.

[52] Jagadish Nayak, Praveena S. Bhat, Udyavara R. Acharya, Choo M. Lim, and Manjunath Kagathi. Automated identification of diabetic retinopathy stages using digital fundus images. *Journal of Medical Systems*, 32(2):107–115, 2008.

[53] NHS Diabetic Eye Screening Programme. Online, May 2013. `http://diabeticeye.screening.nhs.uk`.

[54] Meindert Niemeijer, Michael D. Abràmoff, and Bram van Ginneken. Image structure clustering for image quality verification of color retina images in diabetic retinopathy screening. *Medical Image Analysis*, 10(6):888–898, 2006.

[55] Meindert Niemeijer, Marco Loog, Michael D. Abràmoff, Max A. Viergever, Mathias Prokop, and Bram van Ginneken. On combining computer-aided detection systems. *IEEE Transactions on Medical Imaging*, 30(2):215–223, 2011.

[56] Meindert Niemeijer, Bram van Ginneken, Michael J. Cree, A. Mizutani, G. Quellec, C. I. Sanchez, B. Zhang, R. Hornero, M. Lamard, C. Muramatsu, X. Wu, G. Cazuguel, J. You, A. Mayo, Li Qin, Y. Hatanaka, B. Cochener, C. Roux, F. Karray, M. Garcia, H. Fujita, and Michael D. Abràmoff. Retinopathy online challenge: automatic detection of microaneurysms in digital color fundus photographs. *IEEE Transactions on Medical Imaging*, 29(1):185–195, 2010.

[57] Meindert Niemeijer, Bram van Ginneken, Stephen R. Russell, Maria S. A. Suttorp-Schulten, and Michael D. Abràmoff. Automated detection and differentiation of drusen, exudates, and cotton-wool spots in digital color fundus photographs for diabetic retinopathy diagnosis. *Investigative Ophthalmology & Visual Science*, 48(5):2260–2267, 2007.

[58] João Paulo Papa and Anderson Rocha. Image categorization through optimum path forest and visual words. In *IEEE Intl. Conference on Image Processing*, pages 3586–3589, 2011.

[59] Niall Patton, Tariq M. Aslam, Thomas MacGillivray, Ian J. Deary, Baljean Dhillon, Robert H. Eikelboom, Kanagasingam Yogesan, and Ian J. Constable. Retinal image analysis: concepts, applications and potential. *Progress in retinal and eye research*, 25(1):99–127, January 2006.

[60] Tünde Peto and Christine Tadros. Screening for diabetic retinopathy and diabetic macular edema in the United Kingdom. *Current Diabetes Reports*, 12(4):338–345, 2012.

[61] David J. Pettitt, Wollitzer A. Okada, Jovanovic Lois, Guozhong He, and Ipp Eli. Decreasing the risk of diabetic retinopathy in a study of case management : The california medical type 2 diabetes study. *Diabetes Care*, 28(12):2819–2822, 2005.

[62] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In *IEEE Intl. Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[63] Sam Philip, Alan D. Fleming, Keith A. Goatman, S. Fonseca, P. Mcnamee, Graham S. Scotland, Gordon J. Prescott, Peter F. Sharp, and John A. Olson. The efficacy of automated "disease/no disease" grading for diabetic retinopathy in a systematic screening programme. *British Journal of Ophthalmology*, 91(11):1512–1517, 2007.

[64] Pat J. Phillips. Visible manifestations of diabetic retinopathy. *Medicine Today*, 5(05):83, 2012.

[65] Ramon Pires, Herbert F. Jelinek, Jacques Wainer, and Anderson Rocha. Retinal image quality analysis for automatic diabetic retinopathy detection. In *Intl. Conference on Graphics, Patterns and Images*, pages 229–236, 2012.

[66] Stephen M. Pizer, E. Philip Amburn, John D. Austin, Robert Cromartie, Ari Geselowitz, Trey Greer, Bart Ter Haar Romeny, and John B. Zimmerman. Adaptive histogram equalization and its variations. *Computer Vision, Graphics, and Image Processing*, 39(3):355–368, 1987.

[67] Moacir P. Ponti. Combining classifiers: from the creation of ensembles to the decision fusion. In *Intl. Conference on Graphics, Patterns and Images*, pages 1–10. IEEE, 2011.

[68] Frédéric Precioso and Matthieu Cord. Machine learning approaches for visual information retrieval. In *Visual Indexing and Retrieval*, SpringerBriefs in Computer Science, pages 21–40. Springer New York, 2012.

[69] Gwénolé Quellec, Mathieu Lamard, Michael D Abràmoff, Etienne Decencière, Bruno Lay, Ali Erginay, Béatrice Cochener, and Guy Cazuguel. A multiple-instance learning

framework for diabetic retinopathy screening. *Medical Image Analysis*, pages 1228–1240, 2012.

[70] Gwénolé Quellec, Mathieu Lamard, Pierre M. Josselin, Guy Cazuguel, Béatrice Cochener, and Christian H. Roux. Optimal wavelet transform for the detection of microaneurysms in retina photographs. *IEEE Transactions on Medical Imaging*, 27(9):1230–1241, 2008.

[71] Mansoor Raza, Iqbal Gondal, David Green, and Ross L. Coppel. Classifier fusion using dempster-shafer theory of evidence to predict breast cancer tumors. In *IEEE Region 10 Annual International Conference*, pages 1–4, 2007.

[72] Anderson Rocha, Tiago Carvalho, Herbert F. Jelinek, Siome Goldenstein, and Jacques Wainer. Points of interest and visual dictionaries for automatic retinal lesion detection. *IEEE Transactions on Biomedical Engineering*, 59(8):2244–2253, 2012.

[73] Lior Rokach. *Pattern classification using ensemble methods*, volume 75. World Scientific Publishing Company, 2010.

[74] Jinan B. Saaddine, Amanda A. Honeycutt, K. M. Venkat Narayan, Xinzhi Zhang, Ron Klein, and James P. Boyle. Projection of diabetic retinopathy and other major eye diseases among people with diabetes mellitus: United states, 2005-2050. *Archives of Ophthalmology*, 126(12):1740–1747, 2008.

[75] Walter Scheirer, Anderson Rocha, Ross Micheals, and Terrance Boult. Robust fusion: extreme value theory for recognition score normalization. In *European Conference on Computer Vision*, pages 481–495, 2010.

[76] Chanjira Sinthanayothin, James F. Boyce, Tom H. Williamson, Helen L. Cook, Evelyn Mensah, Shantanu Lal, and David Usher. Automated detection of diabetic retinopathy on digital fundus images. *Diabetic Medicine: a Journal of the British Diabetic Association*, 19(2):105–112, 2002.

[77] Josef Sivic, Bryan C. Russell, Alexei A. Efros, Andrew Zisserman, and William T. Freeman. Discovering objects and their location in images. In *IEEE Intl. Conference on Computer Vision*, volume 1, pages 370–377, 2005.

[78] Josef Sivic and Andrew Zisserman. Video Google: A Text Retrieval Approach to Object Matching in Videos. In *IEEE Intl. Conference on Computer Vision*, pages 1470–1477, 2003.

[79] Joao V. B. Soares, Jorge J. G. Leandro, Roberto M. Cesar, Herbert F. Jelinek, and Michael J. Cree. Retinal vessel segmentation using the 2-d gabor wavelet and supervised classification. *IEEE Transactions on Medical Imaging*, 25(9):1214–1222, 2006.

[80] Akara Sopharak, Bunyarit Uyyanonvara, Sarah Barman, and Thomas H. Williamson. Automatic detection of diabetic retinopathy exudates from non-dilated retinal images using mathematical morphology methods. *Computerized Medical Imaging and Graphics*, 32:720–727, 2008.

[81] Geoffrey K. P. Spurling, Deborah A. Askew, Noel E. Hayman, Naomi Hansar, Anna M. Cooney, and Claire L. Jackson. Retinal photography for diabetic retinopathy screening in indigenous primary health care: the inala experience. *Australian and New Zealand Journal of Public Health*, 34:S30–S33, 2010.

[82] Hugh R. Taylor, Jing Xie, Sarah S. Fox, Ross A. Dunn, Anna Lena M. R. Arnold, and Jill E. Keeffe. The prevalence and causes of vision loss in indigenous australians: the national indigenous eye health survey. *Medical Journal of Australia*, 192(6):312–318, 2010.

[83] Bart Thomee, Mark J. Huiskes, Erwin Bakker, and Michael S. Lew. Large scale image copy detection evaluation. In *Intl. Conference on Multimedia Information Retrieval*, pages 59–66. ACM, 2008.

[84] David Usher, Martin J. Dumskyj, Mitsutoshi Himaga, Tom H. Williamson, Stephen S. Nussey, and James F. Boyce. Automated detection of diabetic retinopathy in digital retinal images: a tool for diabetic retinopathy screening. *Diabetic Medicine*, 21(1):84–90, 2004.

[85] Jan van Gemert, Cor J. Veenman, Arnold W. M. Smeulders, and Jan-Mark Geusebroek. Visual word ambiguity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(7):1271–1283, 2010.

[86] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[87] Peter J. Watkins. ABC of diabetes: retinopathy. *British Medical Journal*, 326(7395):924–926, 2003.

[88] Daniel Welfer, Jacob Scharcanski, and Diane Ruschel Marinho. A coarse-to-fine strategy for automatically detecting exudates in color eye fundus images. *Computerized Medical Imaging and Graphics*, 34(3):228–235, 2010.

[89] John Winn, Antonio Criminisi, and Thomas P. Minka. Object categorization by learned universal visual dictionary. In *IEEE Intl. Conference on Computer Vision*, pages 1800–1807, 2005.

[90] David H. Wolpert. The lack of a priori distinctions between learning algorithms. *Neural Computation*, 8(7):1341–1390, 1996.

[91] World Health Organization. Diabetes programme. Online, May 2013. `http://www.who.int/diabetes/en`.

[92] Dag Wormanns, Karl Ludwig, Florian Beyer, Walter Heindel, and Stefan Diederich. Detection of pulmonary nodules at multirow-detector ct: effectiveness of double reading to improve sensitivity at standard-dose and low-dose chest ct. *European Radiology*, 15(1):14–22, 2005.

[93] Lei Xu, Adam Krzyzak, and Ching Y. Suen. Methods of combining multiple classifiers and their applications to handwriting recognition. *IEEE Transactions on Systems, Man and Cybernetics*, 22(3):418–435, 1992.

[94] Naveed Younis, Deborah M. Broadbent, Simon P. Harding, and Jiten P. Vora. Prevalence of diabetic eye disease in patients entering a systematic primary care-based eye screening programme. *Diabet Med*, 19:1014–1021, 2002.

[95] Wong Li Yun, Udyavara R. Acharya, Yedatore V. Venkatesh, Caroline Chee, Lim Choo Min, and E. Yin Kwee Ng. Identification of different stages of diabetic retinopathy using retinal optical images. *Information Sciences*, 178(1):106–121, 2008.