UNIVERSIDADE DE LISBOA
INSTITUTO SUPERIOR TÉCNICO

Automatic Detection of Melanomas Using Dermoscopy Images

Ana Catarina Fidalgo Barata

Supervisor: Doctor Jorge dos Santos Salvador Marques
Co-supervisor: Doctor M. Emre Celebi

Thesis approved in public session to obtain the PhD Degree in
Electrical and Computer Engineering
Jury final classification: Pass with Distinction and Honour

Jury

Chairperson: Chairman of the IST Scientific Board
Members of the Committee:
Doctor M. Emre Celebi
Doctor Mário Alexandre Teles de Figueiredo
Doctor José Alberto Rosado dos Santos Victor
Doctor Jaime dos Santos Cardoso
Doctor Ana Maria Rodrigues de Sousa Faria de Mendonça
Doctor Alexandre José Malheiro Bernardino

2017

UNIVERSIDADE DE LISBOA
INSTITUTO SUPERIOR TÉCNICO

Automatic Detection of Melanomas Using Dermoscopy Images

Ana Catarina Fidalgo Barata

Supervisor: Doctor Jorge dos Santos Salvador Marques
Co-supervisor: Doctor M. Emre Celebi

Thesis approved in public session to obtain the PhD Degree in
Electrical and Computer Engineering
Jury final classification: Pass with Distinction and Honour

Jury

Chairperson: Chairman of the IST Scientific Board
Members of the Committee:
Doctor M. Emre Celebi, Professor, University of Central Arkansas, USA
Doctor Mário Alexandre Teles de Figueiredo, Professor Catedrático do Instituto Superior Técnico da Universidade de Lisboa
Doctor José Alberto Rosado dos Santos Victor, Professor Catedrático do Instituto Superior Técnico da Universidade de Lisboa
Doctor Jaime dos Santos Cardoso, Professor Associado (com Agregação) da Faculdade de Engenharia da Universidade do Porto
Doctor Ana Maria Rodrigues de Sousa Faria de Mendonça, Professora Associada da Faculdade de Engenharia da Universidade do Porto
Doctor Alexandre José Malheiro Bernardino, Professor Associado do Instituto Superior Técnico da Universidade de Lisboa

FUNDING INSTITUTIONS
FUNDAÇÃO PARA A CIÊNCIA E A TECNOLOGIA

2017

Research is to see what everybody else has seen, and to think what nobody else has thought.

Albert Szent-Györgi (1893-1986) U. S. biochemist

Abstract

Melanoma is one of the deadliest forms of cancer. Unfortunately, its incidence rates have been increasing all over the world. One of the techniques used by dermatologists to diagnose melanomas is an imaging modality called dermoscopy, in which the skin lesion is inspected using a magnification device and a light source. This makes it possible for the dermatologist to observe subcutaneous structures that would be invisible otherwise. However, the use of dermoscopy is not straightforward, requiring years of practice. Moreover, the diagnosis is many times subjective and difficult to reproduce. Therefore, it is necessary to develop automatic methods that will help dermatologists provide more reliable diagnosis.

The goal of this thesis is to develop methods to automatically diagnose dermoscopy images, using image processing and pattern recognition methods. The first phase of this thesis was the development of an algorithm to detect pigment network. This is one of the most relevant dermoscopic structures (according to a dermatologists who collaborated in this work), used for both distinguishing between the two main categories of skin lesions (melanocytic and non-melanocytic lesions), and for diagnosing melanomas. The proposed algorithm uses a bank of directional filters to detect the lines of the network, followed by an analysis of the spatial properties of the network.

The second phase of this thesis focused on the development of computer aided diagnosis (CAD) systems for melanoma diagnosis, using classical pattern recognition approaches. The proposed systems perform very well compared to state-of-the-art systems. The most relevant contributions are: i) a study on the relevance of global features (color, texture, shape, and symmetry), as well as of the best descriptors to represent each of them; ii) the proposal of local features to try to mimic the medical analysis and look for relevant dermoscopic cues; iii) the use of color normalization techniques to make the CAD system independent of the acquisition setup, allowing us to integrate the system in a telemedicine framework; iv) the study of feature fusion techniques.

A clinically inspired CAD system was developed during the third phase of this thesis. Dermatologists do not easily accept ”black box” CAD system that do not provide comprehensive information to justify and validate the automatic diagnosis. This thesis attempted to deal with this problem by proposing a CAD system that bases its decision on features that have a medical meaning. Moreover, the system localizes the medical features within the lesion, making it possible for the dermatologist to validate the output. The main challenge of this approach is the lack of training data with detailed annotations of both the dermoscopic criteria and their corresponding segmentations. The proposed system deals with this problem using a probabilistic image annotation algorithm. This algorithm is trained to detect various dermoscopic criteria using only text labels, without needing detailed segmentations of the criteria. The proposed system is able to justify the diagnosis on medical ground, being trained with a small amount of text labels that are easier to obtain than segmentations. This is the first system of its kinds and achieves a very promising performance.

Keywords

Melanoma detection, dermoscopy, computer aided diagnosis system, feature extraction, image classification, image annotation.

Resumo

O melanoma da pele é um tipo de cancro muito agressivo, cuja incidência tem vindo a aumentar em todo o Mundo. O seu diagnóstico pode ser feito recorrendo a uma técnica de imagiologia designada por dermoscopia, que amplia a lesão e permite observar estruturas subcutâneas. Contudo, a utilização desta técnica requer um longo treino, sendo o diagnóstico muitas vezes subjetivo e difícil de reproduzir. Por estes motivos, torna-se necessário desenvolver métodos automáticos de diagnóstico, que possam ser utilizados pelos dermatologistas.

O trabalho realizado nesta tese tem como objetivo o desenvolvimento de métodos de análise de imagem e de reconhecimento de padrões para diagnosticar imagens de dermoscopia. Este trabalho divide-se em três fases. Numa primeira fase foi proposto um algoritmo para a deteção de rede pigmentar. Esta é uma das estruturas dermoscópicas de referência para os dermatologistas, uma vez que pode ser utilizada para distinguir as duas categorias principais de lesões da pele (melanocíticas e não melanocíticas), assim como para diagnosticar melanomas. O algoritmo baseia-se num banco de filtros direcionais usado na deteção de estruturas lineares, complementado com uma análise espacial das propriedades da rede.

Na segunda fase do trabalho desenvolveram-se sistemas computorizados de apoio ao diagnóstico (CAD) para detetar melanomas, utilizando modos clássicos de reconhecimento de padrões. Os sistemas desenvolvidos têm um bom desempenho. Destacam-se desta fase as seguintes contribuições originais: i) foi feito um estudo sobre a importância das características globais da lesão (cor, textura, forma e simetria), procurando identificar os melhores descritores que as representam; ii) foi proposto o uso de características locais da imagem como forma de aproximação da análise visual feita pelos dermatologistas na deteção de estruturas relevantes; iii) foram estudadas técnicas para tornar os sistema CAD desenvolvidos independentes do equipamento de aquisição, o que lhes permite operar em redes integradas de cuidados médicos envolvendo diferentes hospitais (telemedicina); iv) foram estudadas técnicas de fusão de informação.

Na terceira fase do trabalho desenvolveu-se um sistema CAD inspirado na prática clínica. A comunidade médica tem dificuldade em aceitar sistemas do tipo ”caixa preta” que não oferecem meios de compreensão e validação do diagnóstico proposto. Para colmatar este facto, desenvolveu-se um sistema CAD baseado na deteção de características clínicas, que possibilita também a sua localização na imagem. Esta abordagem é, geralmente, limitada pela ausência de grandes quantidades de dados de treino com anotações clínicas detalhadas, incluindo a segmentação de cada critério clínico. Para lidar com este problema, propõe-se um sistema que utiliza métodos de anotação de imagem baseados em modelos probabilísticos de variáveis latentes. O objetivo é detetar diferentes critérios clínicos a partir de anotações de texto, dispensando, portanto, a segmentação detalhada dos mesmos nas imagens de treino. O sistema proposto oferece uma explicação da decisão tomada e é treinado com recurso a um conjunto modesto de anotações de texto, fáceis de obter. Este sistema obtém um bom desempenho, sendo o primeiro a cumprir os objectivos acima enunciados.

Palavras Chave

Deteção de melanomas, dermoscopia, sistemas de apoio ao diagnóstico, extração de características, classificação de imagem, anotação de imagem.

Acknowledgments

This PhD thesis was supported by FCT in the scope of grant SFRH/BD/84658/2012 and projects PTDC/SAUBEB/103471/2008, [UID/EEA/50009/2013], and PTDC/EEIPRO/0426/2014.

I have always compared my PhD to a journey, where one starts full of expectations but little knowledge about what lays ahead. Doing a PhD was something that I desired since my first day at college, mainly due to my romantic ideas about research. Now that this journey is coming to an end, I can say with all certainty that doing a PhD was the right choice. I was (and hopefully will always be) lucky enough to do something that I truly enjoy: do a lot of research, come up with new ideas, and do not dare to give up until they work. For all of this I am grateful.

Like in any journey, one meets several people along the way. The first that I will name is the one that gets one of my most heartfelt acknowledgments: Professor Jorge Salvador Marques. Thank you so much for trusting and accepting to work with me since the beginning, even when I was a just a Master student finishing her degree, without almost any knowledge of both image processing and machine learning. Thank you for being a supportive supervisor that first guided me with ideas and suggestions, and later on allowed me to follow my on path and try new, inventive, and, sometimes crazy, things. Any of my ideas was always followed by great brainstorming meetings. For me, this is what a supervisor should be all about. Finally, thank you for trusting that I will deliver everything on schedule, even when I try to avoid writing the last paragraph of the thesis because writing is not on the top of my preferences. Thank you for believing in me.

I would like to write another heartfelt thank you to Professor M. Emre Celebi, who accepted to work with me without no other references besides a couple of papers. Thank you for your confidence and for sharing with me your knowledge about dermoscopy image analysis, which allowed me to do some of things in this thesis that would be inaccessible otherwise. Thank you for being accessible all the time. Finally, thank you for the way you received me at your university. I will never forget how you and your family opened your house and friendship to me.

Other people have also played an important role during my PhD. Namely Dr. Jorge Rozeira and Dra. Joana Rocha from Hospital Pedro Hispano, who shared with me their medical knowledge regarding skin lesions and dermoscopy; Prof. Teresa Mendonça from Faculdade de Ciências - Universidade do Porto, who was one of the minds behind the PH^2 dataset and the ADDI project, allowing me to have access to my first dataset of dermoscopy images; and Prof. Mário Figueiredo from Instituto the Telecomunicações - Instituto Superio Técnico, who kindly worked with me into one of the chapters of this thesis. Although she doesn't work in research, I would also like to thank Farida Hamza for all of her kindness during my stay in the US. Your friendship and introducing my to yoga (that I now love) have made wonders for me and have positively contributed to my work.

Life is more than work, thus I feel the need to thank all of my friends that made everything easier. Thank you for all the lunches, dinners, movie sessions, board game nights, and croissant moments! I was trying to avoid naming everyone just because I don't want to forget any name, but there are a couple of people that I have to mention: Rodolfo Abreu and Joana Pinto. Both have shared this journey with me since the beginning and deserve a special word of acknowledgment. Thank you for being crazy, funny, and such good friends. Joana, thank you for remaining happy and being such a nice person, even when my temper gets the best of me and I start treating everyone poorly. Rodolfo, my small dictator, thank you for not even trying to bully me around and for always making the arrangements for my birthdays. One of the best things of this lab is having such good friends around the entire day. Another friend that I would like to remember is Filipe Condessa, who is an old friend and shared a lot of lunches with me. A lot of those lunches resulted in nice ideas that could be applied to my work, so thank you.

Having hobbies is a nice way to get your head out of work, especially when something is not working. Thus, I would like to thank my wind band for all the rehearsals and concerts that make me so happy. Even when I was tired or had just arrived from a conference on the other side of the world, I would still run to the rehearsals, just to enjoy some relaxing moments. I would also like to thank my tango friends (I won't name them, because I don't want to forget anyone), for all the funny classes and milonga nights. Tango and the PhD started at almost the same time, and I can't avoid making a connection.

To my parents Carlos and Lídia, who have stoically survived all of those people saying ”When will your daughter find a job and quit studying?”. Thank you for giving me the wings to fly and do what I like! Thank you for surviving my stay in the US. I only stayed for a couple of months, but I know that both of you were dying on the inside with worry. I don't need to say anything more, both of you know how I feel. To my grandfather Américo, who is convinced that one day I will become a professor, showing a confidence that I don't even have. And last, but not least, to my sister Mafalda, who has a very strong opinion about the importance of degrees, but that I know is bursting with proud at the same time.

My last acknowledgment goes to Cajó, aka Carlos, my partner in crime and in life. I could have done this journey alone, but it felt so much better doing it with you... Thank you for not minding to discuss work on the way home, thank you for reading my papers, thank you for listening to my ideas, thank you for sending away my impostor syndrome, whenever it tried to came out, thank you for not being afraid to say that one of my ideas did not make sense of that my paper was difficult to read, even when I would throw a tantrum. Above all, thank you for being here and for being who you are.

Contents

List of Figures

List of Tables

Acronyms

ACC	Accuracy
ANN	Artificial neural network
BCC	Basal cell carcinoma
BoF	Bag-of-features
CAD	Computer aided diagnosis
CD	Correct detection
CEM	Component-wise expectation maximization
CMRM	Cross media relevance model
corr-LDA	Correspondence-latent Dirichlet allocation
CRM	Continuous relevance model
DF	Detection failure
EM	Expectation maximization
FA	False alarme
FN	False negatives
FP	False positives
GLCM	Gray level co-occurrence matrix
GMCC	Gretag-Macbeth Color Checkers
HSV	Hue saturation value
kNN	k-nearest neighbor
LR	Logistic regression
MBRM	Multiple-Bernoulli relevance model
MIML	Multi-instance multi label
MML	Minimum message length
MR	Median rule
PDE	Partial Differential Equation

PH^2	Pedro Hispano Hospital
Pre	Precision
RBF	Radial basis function
Re	Recall
RGB	Red green blue
ROC	Receiver operator characteristics
SE	Sensitivity
SLIC	Simple linear iterative clustering
SP	Specificity
SVM	Support vector machines
TDS	Total dermoscopy score
TN	True negatives
TP	True positives
WHO	World health organization

1

Introduction

Contents

1.1 Motivation

According to the world health organization (WHO) cancer is one of the leading causes of death worldwide, being the second most important cause of death and morbidity in Europe. WHO estimates that the number of people diagnosed with cancer will double in the next two decades [66]. Since cancer mortality can be significantly reduced if the cases are detected and treated during their early stages, it is of major importance to invest research effort in the development of early cancer detection strategies.

Melanoma is the deadliest form of skin cancer due to its high potential to metastasize. It ranks in the ninth position among the most common types of cancer in Europe [66], and approximately 700 new cases are diagnosed each year in Portugal [1]. Following the general trend, the incidence rates of melanoma have been steadily increasing for the past 30 years. Similarly to other cancers, the survival rate of melanomas in advanced stage is low, but an early diagnosed melanoma can be cured by a simple excision [9]. These statistics are the reason why so much effort has been put both in the development of new imaging techniques, that allow a better visualization of skin lesions, and of automatic means to diagnose melanomas.

Dermoscopy is one of the most popular imaging techniques used by dermatologists. It simultaneously allows the magnification of skin lesions and the observation of surface and subsurface structures, which can not be seen by the naked eye [9,102]. Different studies have shown that dermoscopy significantly increases the diagnosis accuracy of dermatologists (*e.g.*, [125]). Nonetheless, this technique can only be efficiently used by trained physicians, and the diagnosis is usually subjective and difficult to reproduce, since it relies on the visual acuity of the practitioner [197]. These drawbacks fostered the research of computer aided diagnosis (CAD) systems that can act as a second opinion tool and be used by non-experienced dermatologists.

Most of the CAD systems found in literature follow a set of sequential steps: artifact removal, lesion segmentation, feature extraction and lesion binary classification (benign or malignant) [103, 139]. These systems achieve good performances in experimental conditions, but usually extract features that do not have medical meaning or that are not easy to explain to physicians [73]. In addition, the feature extraction procedures are often insufficiently described [38], preventing the implementation of the algorithms by other users. Alternatively, it is possible to develop systems that focus on the extraction of relevant dermoscopic features, which are the grounds of medical diagnosis. Despite their proximity with the medical diagnosis and consequent clinical relevance, these systems have been less explored in literature. Moreover, the proposed methods attempt to detect only one or two dermoscopic criteria, and usually do not try to diagnose the lesions based on the clinical information. This thesis proposes new contributions for each of these two kinds of CAD systems.

The work presented in this thesis was done at the Institute for Systems and Robotics, Instituto Superior Técnico. Part of the described work was done in collaboration with the Faculty of Science of the University of Porto and the Dermatology Service of Hospital Pedro Hispano, as a member of the

[1] Source: Liga Portuguesa Contra o Cancro http://ligacontracancro.pt/

ADDI Project[2]. It also benefited from a collaboration with the Louisiana State University in Shreveport, as well as a two month internship at the same university.

1.2 Objectives

This thesis aims to develop machine learning algorithms for the detection of melanomas in dermoscopy images.

Two directions are followed. The first direction tries to solve this problem using standard pattern recognition methods to detect melanomas. These methods consist of extracting features to describe the lesions, followed by the classification using a machine learning algorithm. This procedure is based on abstract image features and does not try to mimic the analyses of medical experts. In such systems, the output of the classifier does not provide useful cues to explain the final decision (melanoma or benign). The second direction aims at developing solutions capable of extracting clinically relevant information. This information is of major importance, since it will help dermatologists understand and interpret the decisions of the CAD system. However, clinically inspired approaches are much harder to implement, as will be shown in this thesis.

The development of strategies for melanoma diagnosis is usually performed assuming that one is working with melanocytic skin lesions. The discrimination between melanocytic and non-melanocytic lesions is still an open problem and few contributions have been proposed to solve it [172]. This thesis also addresses the distinction between the two types of skin lesions (melanocytic and non-melanocytic).

Figure 1.1 shows the three problems addressed in this thesis, as well as the contributions related with each of them. These problems can be organized in a chronological order as follows: i) discrimination between melanocytic and non-melanocytic lesions, ii) detection of melanomas using pattern recognition methods, and iii) development of a clinically oriented CAD system.

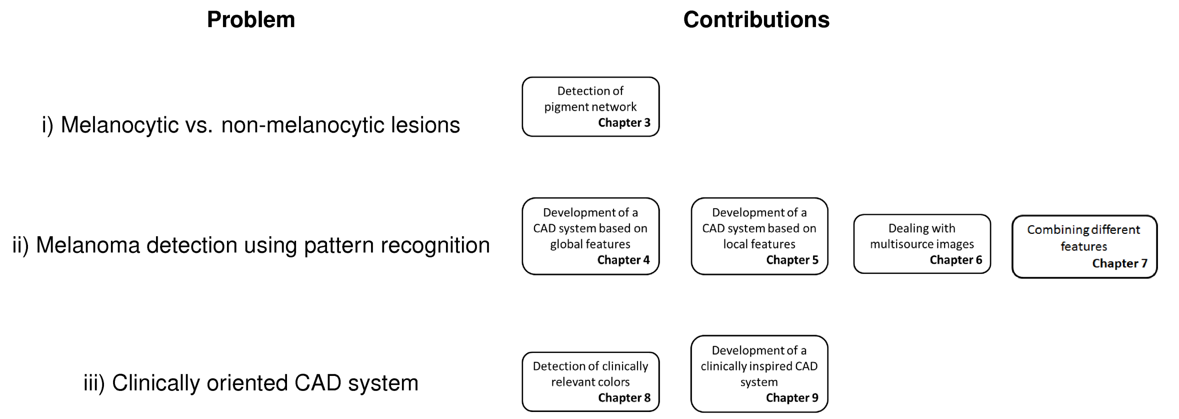

Problem
Contributions
i) Melanocytic vs. non-melanocytic lesions
Detection of pigment network Chapter 3
ii) Melanoma detection using pattern recognition
Development of a CAD system based on global features Chapter 4
Development of a CAD system based on local features Chapter 5
Dealing with multisource images Chapter 6
Combining different features Chapter 7
iii) Clinically oriented CAD system
Detection of clinically relevant colors Chapter 8
Development of a clinically inspired CAD system Chapter 9

Figure 1.1: Thesis problems and contributions.

The *distinction between melanocytic and non-melanocytic lesions* is the first step performed by

[2] Automatic computer-based Diagnosis system for Dermatoscopy Images (PTDC/SAU-BEB/103471/2008)

dermatologists when diagnosing skin lesions. This problem was addressed by developing an algorithm for the detection of pigment network using directional filters.

Melanoma detection was addressed using different strategies. The first approach was the development of a CAD system based on global features (characterizing the entire lesion). In this case, different types of features and classification algorithms were tested in a systematic way. The next step involved the development of CAD systems based on local features (separately characterizing different regions of the lesion), which were implemented using the bag-of-features (BoF) model. As in the case of global features, different features and classifiers were tested. The following contribution was the proposal of a strategy to deal with multi-source images. In clinical practice it is common to find images that were acquired at several clinical facilities, using different devices and illumination. This variety among images can lead to poor performances. To tackle this issue, a color constancy strategy was proposed, in order to normalize the colors of multi-source images. Another contribution was the investigation of strategies to combine global and local features.

A *clinically oriented CAD system* can be achieved by mimicking the medical diagnosis and looking for the presence of clinically relevant clues in dermoscopy images. Although clinically inspired approaches are more challenging to implement, due to difficulty of detecting subtle medical cues, and usually lead to worse performances, it is undeniable that they possess important advantages. They provide not only the diagnosis but also medical information explaining the classifier's decision. In other words, the detected medical features can be used to make the output of the system more clear to dermatologists. This thesis proposes two methodologies to detect medical features. The first method is an algorithm to detect clinically relevant colors in dermoscopy images using Gaussian mixtures. A major limitation of most clinically inspired methods is that they must be trained using detailed annotated data, namely segmentations of the different medical features. When this information is unavailable, it is very difficult to develop methods to detect the medical criteria. This thesis addresses this issue using an image annotation algorithm that is able to detect several medical cues. This information is then used to develop a CAD system for melanoma diagnosis.

1.3 Contributions

The following contributions are presented in this thesis:

- An algorithm for the detection of pigment network, using a bank of directional filters and connected component analysis.
- Development of a CAD system based on a global characterization of the skin lesions. Different types of features and classifiers are studied and compared using this approach.
- Development of a CAD system based on a local characterization of the skin lesions, using the BoF model. Different image sampling strategies, features and classifiers are tested and compared.
- Application of color constancy to deal with multi-source dermoscopy images.

- Development of a CAD system that combines global and local information, as well as different types of features. Early and late fusion strategies are investigated and compared.
- A method for the detection of clinically relevant colors in dermoscopy images using Gaussian mixture models.
- Proposal of a strategy to detect multiple clinical criteria, using an image annotation framework.
- Development of a clinically inspired system that uses medically inspired features to diagnose the skin lesions.

1.4 List of Publications

The work developed in this thesis was partially published in journals, conferences, and book chapters listed below.

Journal Papers

a C. Barata, J. S. Marques, J. Rozeira. ”A system for the detection of pigment network in dermoscopy images using directional filters.” IEEE Transactions on Biomedical Engineering, 59 (10), pp. 2744-2754, 2012.

b C. Barata, M. Ruela, M. Francisco, T. Mendonça, J.S. Marques, ”Two systems for the detection of melanomas in dermoscopy images using texture and color features.”, IEEE Systems Journal, vol.8, no.3, pp. 965-979, 2014.

c C. Barata, M.E. Celebi, J. S. Marques, ”Improving dermoscopy image classification using color constancy.”, IEEE Journal of Biomedical and Health Informatics, vol.19, no.3, pp. 1146-1152, 2015.

d C. Barata, M.E. Celebi, J. S. Marques, ”Clinically inspired analysis of dermoscopy images using a generative model”, Computer Vision and Image Understanding, Elsevier, vol.151, pp. 124-137, 2016.

e C. Barata, M.E. Celebi, J. S. Marques, ”Development of a clinically oriented system for melanoma diagnosis”, submitted to Pattern Recognition, Elsevier.

Book Chapters

f C. Barata, M. Ruela, T. Mendonça, J.S. Marques. ”A bag-of-features approach for the classification of melanomas in dermoscopy images: the role of color and texture descriptors.”, Computer Vision Techniques for the Diagnosis of Skin Cancer, Eds. J. Scharcanski and M. E. Celebi, Springer, pp. 49-69, 2014.

g C. Barata, M.E. Celebi, J.S. Marques, ”Towards a robust analysis of dermoscopy images acquired under different conditions.”, Dermoscopy Image Analysis, Eds. M.E. Celebi, T. Mendonça and J. S. Marques, CRC Press, pp. 1-22, 2015.

Conference Papers

h C. Barata, J.S. Marques, J. Rozeira, ”Detecting the pigment network in dermoscopy images: A directional approach.”, Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 5120-5123, 2011.

i C. Barata, J.S. Marques, J. Rozeira, ”A system for the automatic detection of pigment network,” 9th IEEE International Symposium on Biomedical Imaging (ISBI), pp.1651-1654, 2012.

j C. Barata, J.S. Marques, J. Rozeira, ”The role of keypoint sampling on the classification of melanomas in dermoscopy images using bag-of-features.”, Iberian Conference on Pattern Recognition and Image Analysis (IbPRia), pp. 715-723, 2013.

k C. Barata, J.S. Marques, T. Mendonça. ”Bag-of-features classification model for the diagnose of melanoma in dermoscopy images using color and texture descriptors.”, 10th International Conference on Image Analysis and Recognition (ICIAR), pp. 547-555, 2013.

l C. Barata, J.S. Marques, J. Rozeira, ”Evaluation of color based keypoints and features for the classification of melanomas using the bag-of-features model.”, 9th International Symposium on Visual Computing (ISVC), pp. 40-49, 2013.

m C. Barata, J.S. Marques, M. E. Celebi, ”Towards an automatic bag-of-features model for the classification of dermoscopy images: The influence of segmentation.”, 8th International Symposium on Image and Signal Processing and Analysis (ISPA), pp. 274-279, 2013.

n C. Barata, M.A.T. Figueiredo, M.E. Celebi, J.S. Marques, ”Color identification in dermoscopy images using Gaussian mixture models.”, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3611-3615, 2014.

o C. Barata, J.S. Marques, M.E. Celebi. ”Improving dermosocopy image analysis using color constancy.”, IEEE International Conference on Image Processing (ICIP), pp. 3527-3531, 2014.

p C. Barata, M.E. Celebi, J.S. Marques. ”Color detection in dermoscopy images based on scarce annotations.”, Iberian Conference on Pattern Recognition and Image Analysis (IbPRia), pp. 309-316, 2015.

q C. Barata, M.E. Celebi, J.S. Marques. ”A clinically oriented system for melanoma diagnosis using a color representation.”, 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 7462-7465, 2015.

r C. Barata, M.E. Celebi, J.S. Marques. ”Melanoma detection algorithm based on feature fusion’, 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 2653-2656, 2015.

s C. Barata, M.E. Celebi, J.S. Marques. ”Analysis of dermoscopy images using weakly annotated data: extraction of multiple criteria’, accepted in IEEE International Symposium on Biomedical Imaging (ISBI), 2017.

Co-authored Works

t J.S. Marques, C. Barata, T. Mendonça, ”On the role of texture and color in the classification of dermoscopy images,” 2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 4402-4405, 2012.

u M. Ruela, C. Barata, T. Mendonça, and J.S. Marques, ”What is the role of color in dermoscopy analysis?”, Iberian Conference on Pattern Recognition and Image Analysis (IbPRia), pp. 819-826, 2013.

v M. Ruela, C. Barata, J.S. Marques,”What is the role of color symmetry in the detection of melanomas?”, 9th International Symposium on Visual Computing (ISVC), pp. 1-10, 2013.

w M. Ruela, C. Barata, J.S. Marques, J. Rozeira, ”A system for the detection of melanomas in dermoscopy images using shape and symmetry features”, to appear in Computer Methods in Biomechanics and Biomedical Engineering, 2015.

x T. Mendonça, P. M. Ferreira, A. R. S. Marçal, C. Barata, J. S. Marques, J. Rocha, and J. Rozeira, ”PH^2 - A public database for the analysis of dermoscopic images.”, Dermoscopy Image Analysis, Eds. M.E. Celebi, T. Mendonça and J. S. Marques, CRC Press, pp. 419-439, 2015.

The association between these publications and the thesis chapters is the following:

- **Chapter 3 - Detection of pigment network:** [a], [h], and [i].
- **Chapter 4 - A CAD system based on global features:** [b], [t], [u], [v], and [w].
- **Chapter 5 - The role of local features:** [b], [f], [j], [k], [l], and [m].
- **Chapter 6 - Analysis of images from multiple sources/hospitals:** [c], [g], and [o].
- **Chapter 7 - Feature fusion:** [r].
- **Chapter 8 - Color detection using Gaussian mixture models:** [n].
- **Chapter 9 - Towards the development of a clinically inspired system:** [d], [e], [p], [q], and [s].

According to Google Scholar® the number of citations of all the papers in February 2017 is 255. The three most cited papers are [b] (73), [a] (49), and [n] (18).

2

Dermoscopy Image Analysis

Contents

This section provides basic information about the problem addressed in this thesis, as well as an overview of different computer aided diagnosis (CAD) systems. It can be skipped if the reader is acquainted with the problem.

2.1 Skin Lesions

There is a wide variety of skin lesions that can be organized in a hierarchical way [1][1]. Figure 2.1 summarizes the hierarchical division of skin lesions. First each lesion is split according to its origin, *i.e.*, the type of skin cells responsible for its genesis. Melanocytic lesions, like melanoma, develop from melanocytes, which are the skin cells responsible for the production of a protein pigment called melanin. Non-melanocytic lesions have origin in other types of skin cells, such as the basal or the squamous cells. The distinction between the two types of lesions is visually performed, based on the presence of absence of a set of dermosocopic features, such as pigment network. The next step involves the identification of the lesion type, *i.e.*, if it is a malignant neoplasm or a benign lesion. If the lesion belongs to the latter group, it is then classified into one of the different types of skin lesions. These decisions are also performed based on a set of dermoscopic characteristics. Figure 2.2 shows examples of the different types of skin lesions.

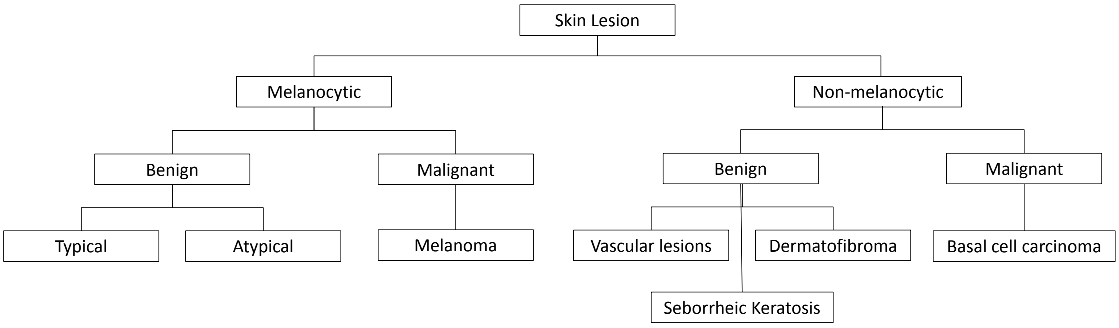


Figure 2.1: Skin lesions classification tree [9].

Non-melanocytic lesions can be divided into benign and malignant neoplasms. Examples of the former are seborrheic keratosis, vascular lesions and dermatofibroma. The malignant neoplasm is called basal cell carcinoma (BCC). This is the most common type of skin cancer, but due to its slow development it is considered less dangerous than melanoma [1].

Melanoma is the malignant form of melanocytic lesions. This lesion grows faster than BCC, showing a high capacity to invade tissues and metastatize to other organs. Its is desirable to detect melanoma in an early stage, when it is still located in the epidermis. This stage is called *melanoma in situ*. If the melanoma is still contained within the epidermis, it has no contact with the deeper layers of the skin and with the vascular plexus (blood flow). This means that it has not yet metastasized and that a simple excision is enough to entirely remove it. Early stage melanomas usually show an irregular shape and several colors, as well as outlined macules or slightly elevated plaques. On the other

[1] A good introduction to dermoscopy can be found in [9].

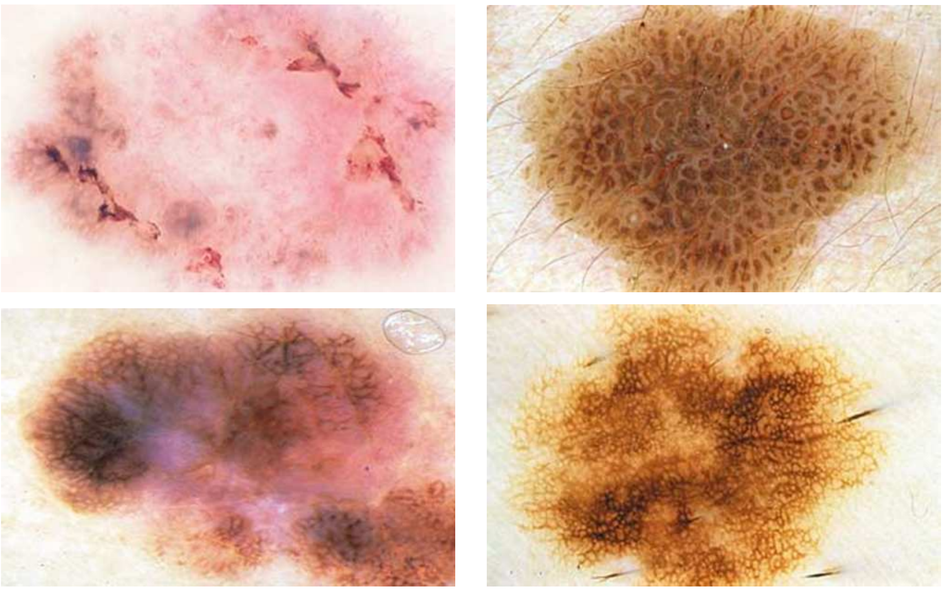


Figure 2.2: Examples of non-melanocytic (1st row) and melanocytic (2nd row) lesions. The lesions are a BCC (1st row, 1st column), a seborrheic keratosis (1st row, 2nd column), and melanoma (2nd row, 1st column), and a benign nevi (2nd row, 2nd column) [9].

hand, invasive melanomas can be papular or nodular, ulcerated and with a coloration that ranges from brown to black, showing also regions of red, white or blue [1].

The remaining melanocytic lesions are divided into typical and atypical lesions. Typical pigmented skin lesions can be acquired or appear at birth (congenital nevi). The genesis of typical lesions occurs in the different layers of the skin, namely on the dermis (deeper layer of the skin), the dermoepidermal junction or on both (compound nevus). Depending on its origin, the lesion will exhibit a specific coloration as well as certain dermoscopic structures [2].

Atypical nevi are irregular pigmented skin lesions. They can be distinguished from benign nevi since they are usually larger, have an irregular and indistinct border, exhibit a mixture of colors and an irregular shape. This type of lesions can be precursors to malignant melanoma. Therefore, it is important to monitor them for any changes in size and color [2].

Despite looking simple, the distinction between melanomas and other skin lesions is not an easy task. This malignant lesion can be incorrectly diagnosed as another type of lesion, which leads to a life threatening false negative. This happens because melanoma often mimics the appearance of other skin lesions, even of lesions that are not melanocytic [9]. In case of uncertainty, dermatologists usually decide to perform and histological exam, which is the only way to identify melanomas without doubt. However, an histological exam usually causes a permanent scar, which is very unpleasant. Moreover, histological exams are expensive and are not suitable to be performed when the patient has too many lesion. CAD systems can be used to overcome this problem, since they work as a second opinion tool [85].

2.2 Medical Diagnosis

This section summarizes the image acquisition methods as well as the medical procedures used to diagnose melanomas.

2.2.1 Skin Lesion Imaging Techniques - Dermoscopy

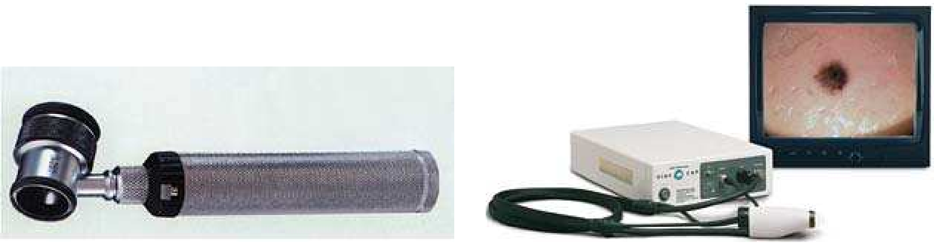
Several image acquisition techniques have been reported for skin lesion inspection. A very common method used for lesion documentation and follow up are clinical images. These images are usually acquired using commercial digital cameras and reproduce what the clinician sees with the naked eye [53]. However, this method renders images with poor resolution and does not allow a deeper inspection of the lesion, which is useful to measure the progress of the disease [115, 175]. Other imaging methods, such as computerize tomography (CT) [178], positron emission tomography (PET) using fludeoxyglucose [144], and magnetic resonance imaging (MRI) [48] are also used to diagnose skin lesions. However, these methods are more suitable to stage the neoplasms and to assess the degree of tissue invasion.

As stated in Section 2.1 it is desirable to diagnose melanoma in its early stage, while it is still *in situ*. Among the aforementioned techniques, the most appropriate for this task is the acquisition of clinical images. However, due to its poor resolution it might not provide sufficient information for a correct diagnosis. Dermoscopy or epuliminescence microscopy was proposed to tackle this drawback. This is a non-invasive inspection technique that allows the visualization of a variety of patterns and structures in skin lesions, which are not discernible to the naked eye [9].

The first step of dermoscopy is to place mineral oil, alcohol or even water on the surface of the lesion. These fluids make epidermis more transparent to light and eliminate part of the surface reflection. After placing the fluid, the lesion is inspected using a dermatoscope, a stereomicroscope, a camera or a digital imaging system [9]. Depending on the instrument used, the magnification of a given lesion ranges from 6x to 40x and even up to 100x. With the magnification, pigmented structures within the epidermis, dermoepidermal junction and superficial dermis become visible. These structures are then used to not only distinguish between melanocytic and non-melanocytic lesions, but also to diagnose the lesions as benign or malignant (recall Figure 2.1). Nowadays there are different dermoscopy devices, which render different degrees of magnification. The traditional dermatoscope is shown in Figure 2.3a). Some health facilities are also equipped with a digital imaging system (see Figure 2.3b)), which allows the acquisition of digital images (see the examples of Figure 2.4).

Several studies have demonstrated that dermoscopy analysis improves the accuracy of melanoma detection by dermatologists by 10-27%, when compared with a naked eye analysis [125]. However, this technique's performance is highly dependent on the expertise of the dermatologist and even trained users can be lured in some particular cases of melanomas that have no characteristic dermoscopy findings [9].

Figure 2.4 illustrates different examples of dermoscopy images. The reader is now asked to try to diagnose the lesions based on their looks and colors. After a first guess, please turn the page and



a) Dermatoscope b) Videodermatoscope

Figure 2.3: Examples of inspection devices [9].

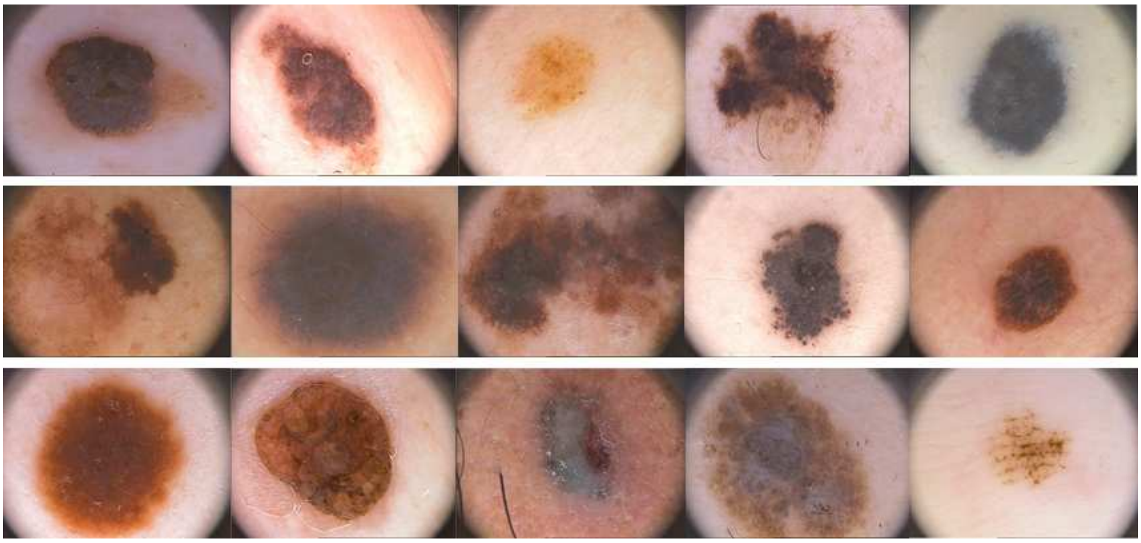


Figure 2.4: Dermoscopy images of melanocytic lesions. Try to guess the diagnose and then see the results in the following page.

see Figure 2.5. It is quite probable that not all your guesses were correct. A lesion might look ”bad” and dangerous and be a benign nevi, while other types of lesions can show a regular aspect and be melanomas.

2.2.2 Melanoma Diagnosis - Medical Procedures

Dermatologists use certain criteria to distinguish between melanocytic and non-melanocytic lesions, and to perform the final diagnosis (benign or malignant). Some of these criteria are related with the set of dermoscopic structures and colors that are observed in the dermoscopy image of the lesion

The first method proposed by dermatologists to diagnose skin lesions is called **pattern analysis** (proposed in 1987 by Pehamberger et al. [141]). This method assumes that there are a set of patterns, also called global features, that can be found in each type of skin lesions. A specific pattern is characterized by one or more dermoscopic structures (local features) that can cover parts or the entire lesion. The patterns considered and their usual aspects are the following:

- **Reticular pattern** - characterized by a pigment network that covers most parts of the lesion. This is the most common pattern in melanocytic lesions.

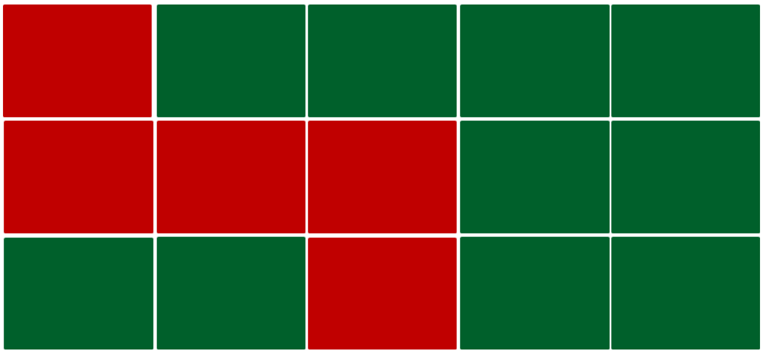


Figure 2.5: Diagnosis of images in Figure 2.4: Red - Melanomas and Green - Benign Lesions.

- **Globular pattern** - characterized by the presence of numerous oval structures, called dots and globules, with different colors and sizes.
- **Cobblestone pattern** - similar to the globular pattern but the globules are closely aggregated.
- **Starburst pattern** - characterized by the existence of pigmented streaks in a radial arrangement localized at the periphery of the lesion.
- **Homogeneous pattern** - diffuse pigmentation, usually brown, blue-gray, gray-black or reddish-black.
- **Parallel pattern** - characterized by pigmented lines that appear in a parallel organization. This pattern is specific of acral regions (palms and soles).
- **Multicomponent pattern** - combination of three or more dermoscopic structures. The presence of this pattern is usually a signal of melanoma.

Both melanocytic and non-melanocytic lesions can be identified and diagnosed using pattern analysis. The final decision as malignant or benign usually depends not only on the number of dermoscopic structures (recall the relation between the multicomponent pattern and melanoma), but also on the shape of the structures that form the pattern. Dermoscopic structures can show a typical or atypical structure, the later being connected with malignant lesions. Pattern analysis has been shown to increase the rate of correct decisions made by dermatologists. However, the assessment of the local dermoscopic structures and consequent pattern characterization is subjective and lacks reproducibility. To tackle this issue more constrained algorithms have been proposed. These algorithms require a prior classification of the lesion as melanocytic or non-melanocytic, which is achieved using Table 2.1 [9].

The **ABCD rule** of dermoscopy (1994) [182] is a procedure that assesses four different characteristics of the lesions: (A)symmetry, (B)order, number of (C)olors, and number of (D)ermoscopic structures. Each of these characteristics is quantitatively scored based on specific criteria (see Table 2.2 for a summary of the criteria). Then, the scores are weighted according to the values of Table 2.2 and a total dermoscopy score (TDS) is computed using (2.1).

$$TDS = 1.3A_{score} + 0.1B_{score} + 0.5C_{score} + 0.5D_{score} \tag{2.1}$$

Table 2.1: Melanocytic algorithm [9]

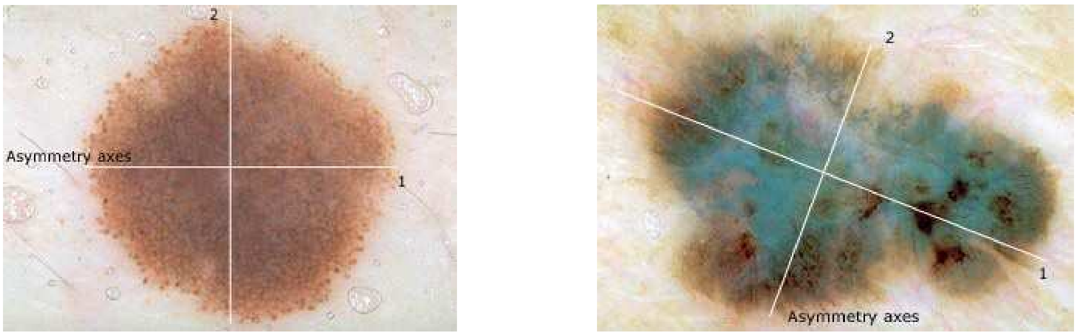
Step	Dermoscopic Criteria	Type of Lesion
I	Pigment network Brown to black dots/globules Streaks Homogeneous blue pigmentation Parallel pattern (palms and soles)	Melanocytic
II	Milia-like cysts Comedo-like openings	Seborrheic keratosis
III	Leaf-like areas Arborizing vessels Irregular gray-blue globules and blotches	Basal cell carcinoma
IV	Red lacunas Red-bluish to red-black homogeneous areas	Vascular lesion
V	Central white patch (surrounded by delicate pigment network)	Dermatofibroma
VI	None of the above criteria	Melanocytic

Table 2.2: ABCD rule of dermoscopy [182]

Criterion	Description	Score	Weight
Asymmetry	In 0,1 or 2 axes; assess not only contour, but also colors and structures.	0-2	1.3
Border	Abrupt ending of pigment pattern at the periphery, in 0-8 segments.	0-8	0.1
Color	Presence of up to 6 colors 1-6 (white, red, light-brown, dark-brown, blue-gray and black).	1-6	0.5
Dermoscopic Structures	Presence of network, structureless or homogeneous areas,streaks, dots and globules.	1-5	0.5

The final diagnosis is made based on the value of the TDS. A lesion is diagnosed as melanoma if the TDS is higher than 5.45 and is considered suspicious if its TDS is in the range 4.8-5.45. Figure 2.6 exemplifies the ABCD rule on two different lesions.

An alternative to the ABCD rule is the **7-point checklist** method (1998) [7]. This procedure requires the identification of 7 atypical local features, usually associated with melanoma. These dermoscopic structures are divided into two classes: major and minor criteria, based on their correlation with a melanoma diagnosis. If any of the criteria is present in the lesion it will receive a score, as show in Table 2.3. In the end, the individual scores are summed up and a total score is computed for the lesion. If the value is above 3 the lesion will be diagnosed as melanoma. Figure 2.7 exemplifies the 7-point Checklist method.



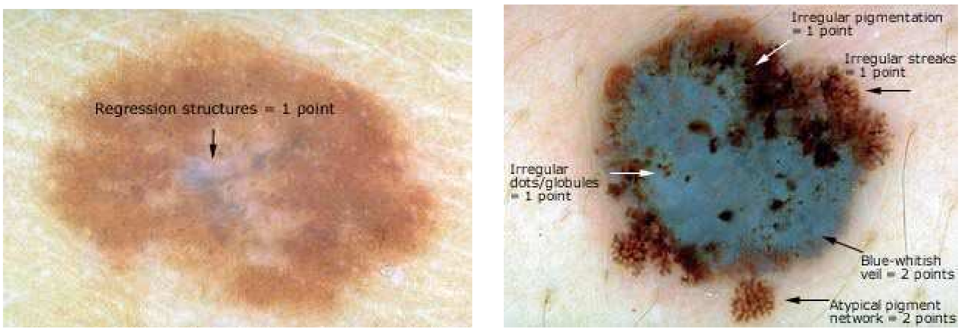
A = 0 x 1.3 = 0;
B = 8x 0.1 = 0.8;
C = 2 [light-brown, dark-brown] x 0.5= 1;
D = 2 [network, globules] x 0.5 = 1;
TDS = 2.8
a) Benign Lesion

A = 2 x 1.3 = 2.6;
B = 5 x 0.1 = 0.5;
4 [light/dark-brown, blue-gray, black, white] x 0.5 = 2;
4 [homogeneous areas, streaks, dots, globules] x 0.5 = 2;
TDS = 7.1
b) Melanoma

Figure 2.6: Examples of the ABCD rule (adapted from [9]).

Table 2.3: 7 point-checklist: criterion and scores [7]

Local Features	Score
Major Criteria	
1.Atypical Pigment Network	2
2.Blue-Whitish Veil	2
3.Atypical Vascular Pattern	2
Minor Criteria	
4.Irregular Streaks	1
5.Irregular pigmentation	1
6.Irregular dots/globules	1
7.Regression structures	1


Regression structures = 1 point
Irregular pigmentation = 1 point
Irregular streaks = 1 point
Irregular dots/globules = 1 point
Blue-whitish veil = 2 points
Atypical pigment network = 2 points

7-point score = 1
a) Benign Lesion

7-point score = 7
b) Melanoma

Figure 2.7: Examples of the 7-point checklist [9].

2.3 CAD Systems

The diagnosis of melanomas using dermoscopy images is a challenging task. Even with the use of medical algorithms such as ABCD rule and 7-point checklist, it is not always easy to discriminate melanoma from other types of skin lesions. This decreases the sensitivity of dermatologists and increases the number of unnecessary histological exams, since this is the only way to correctly diagnose the lesions [9]. A reason for resorting to a histological exam may be related with the inexperience of the dermatologist when working with dermoscopic images [24]. Furthermore, different dermatologists may disagree on their diagnosis of the skin lesions. This happens because the assessment of the different dermoscopic criteria is based on the visual acuity and the experience of the expert.

The previous drawbacks fostered the requirement of automated systems for the diagnosis of melanomas. A CAD system possesses different assets that can be helpful for dermatologists. First, the diagnosis will not depend on the person who is using it at the time, thus the diagnosis will be reproducible. Then, it can be used to follow up a certain lesion. Finally, the system can be used by both experienced and non-experienced dermatologists, working as a second opinion tool or a learning device.

Several CAD systems have been proposed for melanoma detection [103,121,139]. These systems have generic sequence of steps: image preprocessing, lesion segmentation, feature extraction and classification.

i) Preprocessing: Image preprocessing is a required step to deal with images that do not have the optimal quality to be analyzed. This lack of quality can be related with the presence of artifacts that can negatively influence the performance of the subsequent algorithms. The commonly removed artifacts are reflections, skin hairs, and ruler marking. Another important aspect to be considered is color normalization. Dermoscopy images can be acquired using different devices and illumination conditions, rendering non-reliable color information. Therefore, it is important to prevent this issue by including a color correction step in the preprocessing phase.

ii) Lesion segmentation: This is a challenging task that has been thoroughly investigated in literature [36, 40, 174]. The great variety of lesion shapes, sizes, and colors as well as different skin types and textures do not make it easy to develop a robust segmentation algorithm. A correct segmentation is a necessary step to achieve a correct extraction of features and consequent lesion characterization. To prevent misclassification due to incorrect segmentation, several CAD systems are semi-automatic, using manual segmentations performed by experts, instead of automatic ones.

iii) Feature Extraction: Similarly to what dermatologists do to classify skins lesions, in a CAD system it is also necessary to extract information that characterizes the lesions. Several features have already been used to describe skin lesions [121]. The type of features used in this block can be divided into two different classes: classical image analysis features and clinically inspired features [38, 103]. This division leads to the separation of CAD systems in two classes as well: pattern recognition CAD systems and clinically inspired ones. The features used in each case as well as the grounds for their selection will be addressed in the following subsections.

iv) Lesion classification: This is the last step of the CAD system. This task is accomplished by means of classification algorithms. Different algorithms have been tested so far, with a significant preference to artificial neural networks (ANN), support vector machines (SVM), decision trees, and k-nearest neighbor (kNN) [103, 121, 139]. Sometimes an intermediate step of feature selection is performed before lesion classification. The goal of this block is to reduce the size of the feature vector by eliminating irrelevant and/or redundant features.

A main downside of these CAD systems is that most of them were developed to work only with melanocytic lesions and are not able to perform the distinction between melanocytic and non-melanocytic lesions. This suggests that these systems might fail or not be able to classify non-melanocytic lesions. The problem of distinguishing between melanocytic and non-melanocytic lesions has been scarcely addressed in literature. Notable exceptions are the CAD system proposed by Iyatomi et al. [89] that distinguishes between melanocytic and non-melanocytic lesions, and the one proposed in [171], where a four class classification, which includes non melanocytic lesions, is performed.

The following subsections give additional details about the kind of features extracted in each type of CAD system, as well as an overview of the state-of-the-art.

2.3.1 Pattern Recognition CAD systems

The majority of CAD systems found in literature belong to this group. They are usually inspired by the ABCD rule (recall Table 2.2) and try to extract features that can be related to the four criteria of this rule. Thus, the commonly extracted types of features are [103, 121]:

- **Asymmetry features:** metrics are computed to evaluate the degree of asymmetry of the lesion, with respect to its shape, color, and texture.
- **Shape features:** these features try to describe the general shape of the lesion, using attributes such as the area, perimeter, and circularity index. Shape features are associated with the characterization of the lesion's border, thus it is also common to include other features, such as the fractal dimension and the irregularity index.
- **Color features:** describe the color properties inside the lesion. Different features can be used in this case, such as statistical descriptors (mean, standard deviation, skewness, entropy, etc), color histograms, and chromatic differences. Multiple color spaces are usually used to extract the previous features (*e.g.*, RGB, HSV, L*a*b*).
- **Texture features:** the goal of these features is to characterize the dermoscopic structures. Some of the most popular features are the gray level co-occurence matrix (GLCM) [84] and the associated statistics, as well as Gabor filters [10].

Table 2.4 summarizes the performance of a subset of relevant CAD systems. These methods can achieve very promising results. However, dermatologists do not easily accept them, since they do not provide comprehensive clinical information that can be used to understand why the system classifies a lesion as benign or malign [52, 59]. This happens because the used features are expressed in numerical values and it is difficult to correlate them with the dermoscopic criteria used by physicians. Two notorious examples of this situation are the use of abstract texture descriptors to represent dermoscopic structures, instead of detecting the actual structure, and the use of abstract color features to mimic color analysis, while dermatologists count the number of colors. In addition, the data sets used in the studies are not the same and the feature extraction processes are usually poorly described, preventing a fair comparison between different systems.

Table 2.4: Performance of pattern recognition CAD systems. Works identified with ”*” only report accuracy values.

Classification Algorithm	Reference	Sensitivity	Specificity	#Images/#Melanomas
ANN	[151]	94.3%	93.8%	500/200
	[90]	85.9%	86.0%	1258/198
	[129]	67.5%	80.5%	180/72
	[154]	78.4%	95.7%	98/51
SVM	[38]	92.3%	93.3%	564/88
	[148]	72.4%*		358/134
kNN	[31]	98.0%	79.0%	1081/423
	[148]	62.9%*		358/134
	[154]	70.2%	76.5%	98/51
AdaBoost	[128]	92.0%	70.0%	152/42
	[33]	90.0%	77.0%	655/511

2.3.2 Clinically Inspired CAD systems

To overcome the lack of clinical information in CAD systems, some research groups have tried to replace the abstract computer vision features by detectors of dermoscopic criteria. The detected criteria are then used to mimic the dermatologist’s procedure and diagnose the lesion based on clinical grounds. This means that image processing techniques and/or pattern recognition approaches are used to identify the dermoscopic criteria, extract them and, if required, try to assess if they are typical or atypical. Finally, this information is used to classify the lesion as melanoma or benign. However, the development of clinically inspired system is a much harder task and the results achieved so far are very incomplete.

The proposed descriptors are inspired in the diagnosis methods described in Section 2.2.2. Some works focus on the detection of the global patterns (*e.g.*, reticular, globular [9]), using approaches such as texture analysis [183], hidden Markov models [167], and Gaussian mixtures [160]. The detection of the patterns is usually followed by a diagnosis of the lesion as melanoma or benign, as in the pattern analysis method.

Another approach consists of detecting criteria that are related with the ABCD rule, 7-point checklist, or both. The developed methods often focus on color criteria, such as the development of color models to identify clinically relevant colors and quantify them (as in the ABCD rule), or the development of algorithms to detect melanoma related color structures. Examples of the later are the blue-whitish veil [37,56] and regression areas [56], which are used both in the ABCD rule and 7-point checklist. Other commonly detected criteria are pigment network [158], dots [202], and streaks [155], all of which are considered in the ABCD and 7-point checklist. The detection of these structures has to be followed by their characterization as typical or atypical, which is a challenging task.

Table 2.5 summarizes the most relevant works in this field. Clinically inspired CAD systems are significantly more difficult to implement than the ones based on pattern recognition, due to the multitude of criteria that have to be detected and analyzed. In addition they often require a very detailed annotation and segmentation of the images with clinical information, which is very difficult to obtain. The detection of the criteria is an intermediate step, where it is necessary to provide reliable outputs, such that a final diagnosis can be performed. Therefore, several works only focus on the detection of reliable clinical information, without distinguishing between melanoma and benign lesions. In Table 2.5 these works are differentiated from those that propose a final CAD system.

Table 2.5: Clinically inspired CAD systems. Systems identified with an "^{a}" use the detected criteria to diagnose melanomas.

Dermoscopic Criteria	Reference
Global pattern	[127,156,167,183], [3,91,160]^a
Color identification/quantization	[123,142,165], [41,112,161,164]^a
Pigment network	[6,11,70,79,138,158,173,196], [23,57]^a
Dots/globules	[54,70,202]
Streaks	[131,155,157], [57]^a
Blue-whitish veil	[56,119,120], [37,57]^a
Regression structures	[54,56,181], [49,57]^a
Hypopigmentation	[54], [49]^a
Blotches	[117,133,142,180]

Table 2.5 provides relevant information. There is a significant difference in the number of works that deal with global patterns and those that deal with local criteria. It is also clear that there are patterns and structures that are preferred by the research community, namely colors and pigment network. Most of the works attempt to detect only one dermoscopic criteria, which shows that this is a challenging problem. Finally, it is interesting that most of the studies shown in the table refer to the detection of dermoscopic criteria, but do not proceed to the task of melanoma diagnosis. This evidences the difficulty of developing a CAD system based on this type of features.

2.4 Databases

Most of the CAD systems proposed in literature are trained and tested using dermoscopic databases that are acquired at one or more hospitals. Each research group tends to use its own dataset, which differ in size, number of melanomas (recall Table 2.4), and acquisition setups. Moreover, most of these datasets are not publicly available. This does not allow a direct comparison between approaches, which is required to understand the value of each method.

There are some groups who use large commercial datasets that come as support material for dermoscopy atlas (*e.g.*, EDRA [9], which is used in parts of this thesis). The use of these databases reduces the variability between systems, since they are trained using approximately the same set of images. Another main advantage is that commercial databases usually come with medical information, such as the diagnosis and the evaluation of the lesion using the medical criteria described in 2.2.2. However, commercial databases can be expensive and difficult to obtain.

To the best of our knowledge, the first publicly available dataset of dermoscopy images (called PH^2 - Pedro Hispano hospital) was released by the team of ADDI/FCT project in which we participated [2]. Besides the dermoscopy images and their diagnosis, PH^2 provides medical segmentations for the lesions as well as information regarding the presence or absence of relevant medical criteria. Moreover, different groups can download it and compare their the results with the ones obtained using different approaches. PH^2 is used to train and test most of the systems described in this thesis. It is desired that PH^2 can be used for a fair comparison between different systems and that is why it is adopted in most of this work.

Recently a new dataset has been made available for the research community by the international skin imaging collaboration (ISIC). This dataset was released as part of a 2016 conference challenge and contains 900 images for training (173 melanomas) and 379 images for testing (75 melanomas) [82].

2.5 Conclusions

This chapter provided a general overview of the problem addressed in this thesis. It can be divided into two main parts: i) medical knowledge about skin lesions; and ii) an overview overview of the state-of-the-art in dermoscopy image analysis.

The first part started by describing the hierarchical tree that dermatologists follow whenever they diagnose a skin lesion. Then, the methodologies used to inspect skin lesions were presented, with a special emphasis given to dermoscopy, since this is the imaging method investigated in this thesis. Finally, the different methods used by dermatologists to diagnose dermoscopy images were presented: pattern analysis, ABCD rule, and 7-point checklist.

In the second part, the major limitations of dermoscopy were pointed out, namely the subjectivity of the method, its pitfalls, and the need for trained experts. These problems are the grounds for the proposal of CAD systems for melanoma diagnosis. As described in this chapter, CAD systems follow a generic sequence of steps: i) preprocessing; ii) lesion segmentation; iii) feature extraction; and iv) lesion classification. The kind of features extracted during the third step make it possible to distinguish between two classes of CAD systems: i) pattern recognition CAD systems, when the

[2] Project web page http://www.fc.up.pt/addi/.

extracted features are traditional image analysis features; or ii) clinically inspired CAD systems, when the extracted features have a medical meaning and can be associated with dermoscopic cues. Pros and cons have been identified for both systems. Pattern recognition ones achieve promising results, but are not easily accepted by the medical community due to their lack of comprehensible information. Moreover, they are difficult to reproduce due to dataset changes and lack of detailed information regarding the feature extraction process. Clinically inspired systems provide medical information, making them desirable for the medical community, but are much harder to implement. They not only require the detection of a multitude of criteria, but also a detailed annotation of the images with clinical information, which is not easy to obtain. Moreover, few clinically inspired systems actually perform melanoma diagnosis and usually detect only one criterion.

3

Detection of Pigment Network

Contents

3.1 Motivation

Pigment Network is one of the most important dermoscopic structures. Its presence as well as its shape (typical or atypical) are accounted in the process of lesion diagnosis, both in the distinction between melanocytic and non-melanocytic lesions, and in the identification of melanoma [8].

This dermoscopic structured is considered to be the hallmark of melanocytic lesions [9]. To understand the relation between melanocytic lesions and pigment network, it is necessary to know more about its origin. Human skin has two significant layers: the superficial layer called epidermis and the inner layer called dermis. Melanocytes, which are the cells responsible for the production of melanin pigment, can be found in the basal layer of epidermis, next to the dermoepidermal junction.

An analysis of the structure of epidermis shows that it can be compared to a set of connected ridges, separated by valleys. The melanin produced by the melanocytes tends to locate preferentially on the top of the ridges, being closer to the skin surface. Melanin can also be found on the valleys, but in significantly lower concentrations [1]. When a skin lesion is inspected from above, as happens in the case of dermoscopy, the set of connected ridges with high concentrations of melanin seems to form a dark grid pattern, while the valleys with a lower concentration of melanin can be compared to light holes. Combined, they form the structure called pigment network, which can be described as a grid of thin brown lines over a diffuse light brown background. Based on the previous analysis it is possible to understand the relation between pigment network and melanocytic lesions. Figure 3.1 shows two examples of pigment network.

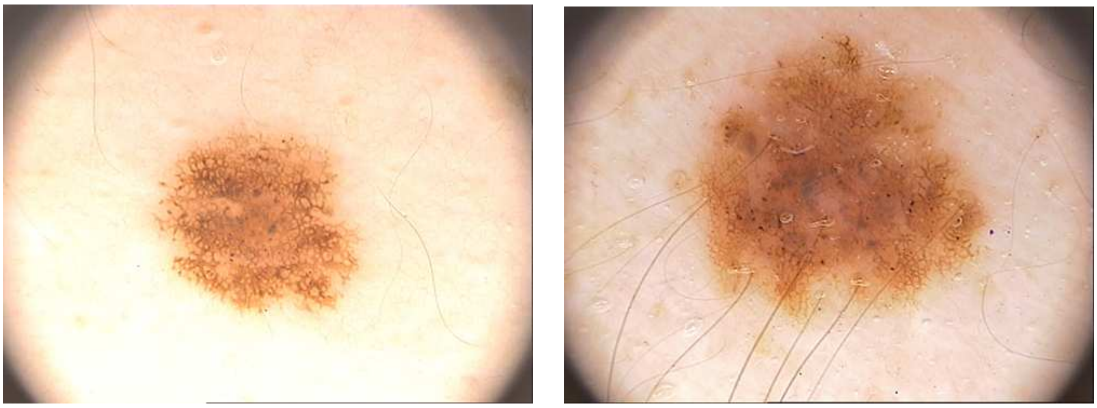


Figure 3.1: Examples of pigment network.

Pigment network is assessed in all the medical procedures (recall Section 2.2.2). During the application of the ABCD rule (Table 2.2) [182] it is necessary to look for the presence of this structure. On 7-point checklist (Table 2.3) [7], pigment network as well as its shape are considered, receiving one of the major scores if it is atypical. Finally, on the pattern analysis method, pigment network is considered as the element of the reticular pattern [141]. The main reason behind the choice of pigment network as a relevant dermoscopic feature is, as before, related with the distribution of melanocytes on the epidermis.

Melanoma results from an abnormal proliferation of mutated melanocytes that can vary in size and amount of produced melanin. These malignant melanocytes have the ability to move freely and to multiply themselves, disrupting the normal structure of epidermis [1]. The abnormal multiplication of melanocytes causes a thickening in some of the ridges of epidermis, consequently leading to a thickening of the grid of pigment network. Furthermore, the increased production of melanin leads to a darkening of the network. Thickened and darker grid lines, as well as an irregular distribution throughout the lesion are the characteristics of an atypical pigment network [9]. Thus, an atypical network is usually a good identifier of melanoma.

It is important to develop strategies not only to detect pigment network, which can help in the automatic decision between melanocytic and non-melanocytic, but also to develop methods to analyze the shape of the network in order to identify the presence of atypical network. The last aspect can be of help in the process of identifying melanoma. The former has been scarcely addressed in literature and it is a missing task in most CAD systems.

3.2 Related Work

One of the first methods to detect pigment network was proposed by Fleming et al. [70]. Their approach consists of applying the first and second derivatives to the gray-scale dermoscopy image in order to detect the pixels that belong to the lines of pigment network. A pixel is considered active if its first derivative is close to zero and the second derivative has a high value. The second derivative is also used to link the pixels and to obtain the final mesh, since it provides information about the orientation and proximity between pixels. A different way to link the pixels was proposed by Grana et al. [79]. Instead of using the second derivative, they apply morphological masks.

Some of the approaches apply a filtering technique to characterize or highlight certain aspects of pigment network. Anantha et al. [6] divide each lesion into small patches and compute the responses of the patches to Laws' energy masks. This information is then used to classify each patch as pigment network or not. Betta et al. [23] combine two different filters to detect pigment network. First, they apply a median filter to a gray level image and remove isolated points using a morphological operation. Then, they apply a high-pass filter in the Fourier domain to exclude any slowly modulating frequencies. The outputs of both filters are combined in the end to obtain a mask that shows the holes of pigment network. Di Leo et al. [57] extended the previous approach to the identification of atypical pigment network. To achieve this goal they compute different features over the holes mask, such as topological information. Sadeghi et al. [158] also detect the holes of pigment network. To highlight the holes they apply a Laplacian of Gaussian filter. Then, they transform the filtered image into a set of graphs and search for those graphs that are cyclic, since these are assumed to be candidates of pigment network. The final identification is performed using a density ratio metric that compares the number of graph edges with its vertices and the area of the whole lesion. An extension of this work was proposed in order to characterize the network as typical or atypical.

Another strategy for the detection of pigment network is the use of supervised machine learning techniques [160, 167, 196]. An example is the work of Wighton et al. [196], where each pixel in the dermoscopic image can be classified as background, when its outside the lesion, present (pigment network) or absent. To describe the pixels they used Gaussian and Laplacian filter sets.

This thesis presents a different approach to detect pigment network using a bank of directional filters. This step provides the user with a mask of the network as well as the identification of its location within the lesion. Then, the lesion is classified as *with* or *without pigment network*, using a trained classifier. Another contribution is the validation of the detected network regions against those identified by an expert[1]. Related works only provide scores for the classification task. Usually, they do not show the location of the detected network, and do not compare it with ground truths provided by the experts. Thus, the output of the algorithm as well as its validation make it unique, when compared with other works.

It is important to stress that the goal of this chapter is to aid in the distinction between melanocytic and non-melanocytic lesions. Thus, the distinction between typical and atypical network is not considered.

3.3 Proposed Method

The proposed method comprises three blocks, as shown in Figure 3.2.

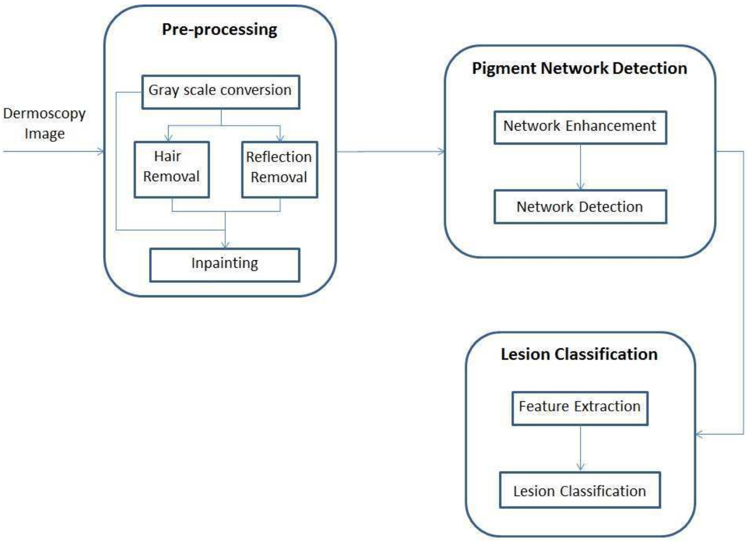


Figure 3.2: Overview of the pigment network detection system.

The first block performs a pre-processing on the given dermoscopy image. The color image is converted into a gray scale image and two types of artifacts are removed in this step: hair and reflections caused by the dermoscopic gel. These artifacts must be removed or attenuated prior to the detection of pigment network since they might occlude part of the reticular pattern or create spurious structures in the output image. Different methodologies have been used to convert the RGB dermoscopy images into gray scale (*e.g.*, [159]). Silveira et al. [174] found out that a simple and efficient strategy to convert dermoscopy images is to select the RGB channel with the highest entropy value. According

[1] The ground truth was kindly provided by Dr. Jorge Rozeira.

to the findings in [174], the highest entropy channel is the one that usually performs best in the lesion segmentation task. Therefore, this approach will be used in this work.

In the second block, regions with pigment network are detected using two of its distinctive properties: intensity (*i.e.*, the transitions between the dark lines and the lighter ”holes”) and geometry or spatial organization (it is assumed that the lines of pigment network form connected structures). The intensity property will be used to perform an enhancement of the network by applying a bank of directional filters, while the spatial organization will be used to perform the actual detection and generation of a binary *net-mask*.

The final block aims to assign a binary label to each image: *with* or *without pigment network*. To accomplish this objective, features which characterize the topology of the detected regions in a given image are extracted and used to train a classifier, namely AdaBoost [72].

Figure 3.3 shows the output of each processing block in the case of a melanocytic skin lesion with pigment network.

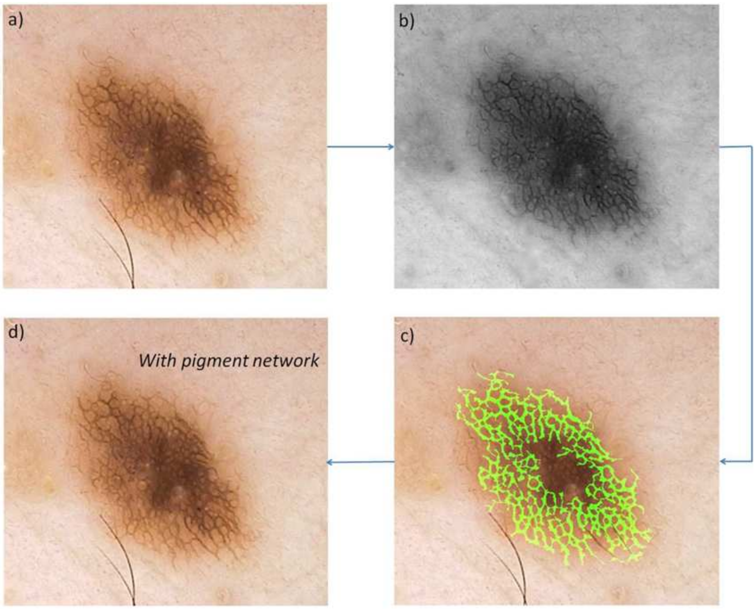


Figure 3.3: Outputs of the proposed system: a) Original image; b) Output of the pre-processing block; c) Output of the pigment network detection block and d) Output of the lesion classification block. Image from [16]

3.4 Directional Filters

The lines of pigment network can be seen as linear strokes, with different orientations. Therefore, directional filters can be used to enhance them. In this thesis, a bank of directional filter was designed for this purpose. These filters are inspired both on the concept of 2D Gabor filters [51], which simulate the behavior of the simple cells of receptive fields (cells that respond to a line or edge of a certain orientation, called directional cells) [97], and on the principle of matched filtering which determines that, in order to detect the presence of a known structure in an image with additive Gaussian noise, the optimal solution consists of filtering the image with a linear filter that has a mask equal to the structure to be detected [189].

Since the lines of pigment network not only have an unknown direction but may appear in a dermoscopy image with several orientations as well, a bank of directional filters is adopted. The filter bank consists of $N+1$ filters, each one of them tuned to a specific orientation $\theta_i \in [0,\pi]$, $i = 0, ..., N$. The impulse response h_{θ_i} of each one has a linear shape, inspired by the lines of pigment network, and is given by

$$h_{\theta_i}(x,y) = G_1(x,y) - G_2(x,y)\,, \tag{3.1}$$

where G_k is a Gaussian filter:

$$G_k(x,y) = C_k \exp\{-\frac{x'^2}{2\sigma_{x_k}^2} - \frac{y'^2}{2\sigma_{y_k}^2}\}, k = 1, 2. \tag{3.2}$$

In (3.2) C_k is a normalization constant and the values of (x', y') are related with (x, y) by a rotation of amplitude θ_i.

$$x' = x\cos\theta_i + y\sin\theta_i \tag{3.3a}$$

$$y' = y\cos\theta_i - x\sin\theta_i\,. \tag{3.3b}$$

The values for the parameters σ_{x_k} and σ_{y_k} are chosen in such a way that the second filter is highly directional and the first one is less directional or even isotropic. A difference of Gaussians was chosen since it allows a good enhancement of directional structures while removing the effect of the background.

The image I is filtered by each directional filter. The output of the $i-th$ directional filter is given by the following convolution

$$I_i(x,y) = h_{\theta_i}(x,y) * I(x,y)\,. \tag{3.4}$$

To combine the output of the N + 1 directional filters, a selection of the maximum output at each pixel (x, y) is performed

$$J(x,y) = \max_i I_i(x,y)\,. \tag{3.5}$$

Figure 3.4 summarizes the directional filters bank.

3.5 Removal of Image Artifacts

The dermoscopy images used in computer-aided diagnosis often display artifacts such as skin hairs or reflections produced during the acquisition process. These artifacts may interfere with many computational procedures required for accurate diagnosis, such as border detection and dermoscopic criteria extraction. Therefore, a preprocessing step in which these artifacts are detected and removed is required in order to ensure that the results obtained by any algorithm are not compromised. The algorithms described in this section will also be used in all the remaining chapters.

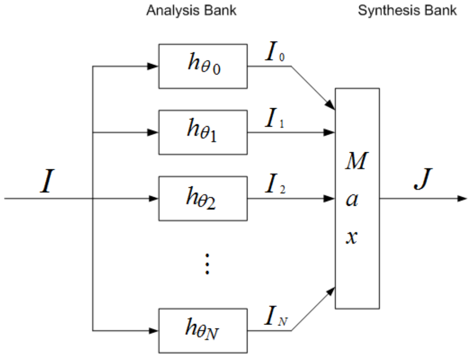


Figure 3.4: Directional Filters Bank.

Reflection Detection: Reflection artifacts appear in dermoscopy images during the acquisition process. The method used for their detection is a simple thresholding algorithm. It is assumed that a pixel (x, y) is classified as a reflection artifact if its intensity is high and higher than the average intensity $I_{avg}(x, y)$ of its neighbors. This statement is summarized in the following condition

$$\{(x, y) : I(x, y) > T_{R1} \wedge I(x, y) - I_{avg}(x, y) > T_{R2}\}, \tag{3.6}$$

where I is the gray level image, I_{avg} is the average intensity in a local neighborhood of the pixel and $T_{R1} = 0.7$, $T_{R2} = 0.098$ are threshold values, which were experimentally obtained. These values were defined for images with intensity in the range $[0, 1]$. I_{avg} is computed using a local mean filter with dimensions 11×11.

Hair Detection: Hair artifacts are highly directional structures with a linear shape that can occlude parts of the lesion. These have a shape similar to a stroke of the pigment network. Therefore, a similar approach can be used to detect both of them. The hair detection algorithm uses a bank of N = 64 directional filters (recall Section 3.4) to perform the detection. The filter parameters were experimentally obtained and are given by: $\sigma_{x_1} = 20$, $\sigma_{y_1} = 6$, $\sigma_{x_2} = 20$, and $\sigma_{y_2} = 0.5$. The mask of the filters have a dimension of 41×41. These values are suitable for images with an average resolution of 573×765. After filtering the gray level image I, a threshold is applied to the output J. If $J(x, y)$ exceeds a threshold $T_H = 0.06$ it is classified as hair.

The gaps that appear after removing the artifacts are filled using a partial differential equation-based interpolation, also known as inpaiting [22]. This operation fills the unknown regions using information of their neighborhood. Examples of preprocessed images can be seen in Figure 3.5.

3.6 Detection of Pigment Network

Pigment network has two main characteristics. First, there is a significant color difference between the darker lines and lighter holes. Second, the lines of the network form a linked structure [9]. These

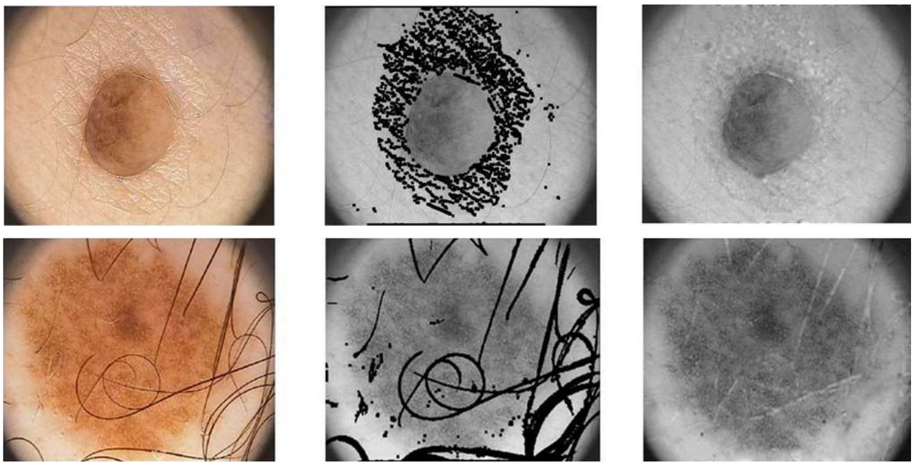


Figure 3.5: Examples of the preprocessing process: original (left); gray level image after artifact detection and removal (mid); output of the preprocessing block, after applying the inpaiting step (right). Images from [16]

two aspects can be used to distinguish pigment network from other dermoscopic structures. The proposed method makes use of both properties, in order to extract a *net-mask* that highlights pigment network regions.

The first step of the method consists of enhancing the lines of pigment network, by increasing the contrast between them and the background. This task is performed using the bank of directional filters described in Section 3.4, with the following empirically determined parameters: $\sigma_{x_1} = 40$, $\sigma_{y_1} = 40$, $\sigma_{x_2} = 3$, and $\sigma_{y_2} = 0.5$. The size of the masks are 11×11 and $N = 9$ directional filters are used. These values were obtained for images with an average resolution of 573×765. Then, the output of the directional filters is thresholded in order to find the candidate pixels that belong to pigment network.

The second step makes use of the geometrical properties of pigment network, namely it is assumed that pigment network consists of a set of connected lines. The connectivity between pixels can be identified using connected component analysis. This technique can be applied with different types of neighborhood. In this thesis, a 8-connectivity criterion was adopted, since it considers information from vertical, horizontal, and diagonal neighbors. This approach ensures the detection of larger regions and the linking of regions that could be missed if the 4-connectivity criterion was used. Then, all the connected components in the binary image are extracted and classified by comparing their areas with a threshold. The regions that fulfill the following criterion are classified as pigment network

$$A(R_c) > A_{min}, \tag{3.7}$$

where R_c is the $c-th$ connected region, $A(R_c)$ is its area and A_{min} is a threshold experimentally set to 900 pixels. By enforcing the previous condition it is possible to discard all the connected components with small areas, which can be seen as noise.

The final pigment network region R is computed as the union of all the connected components which meet condition (3.7)

$$R = \bigcup_{c:A(R_c)>A_{min}} R_c, \tag{3.8}$$

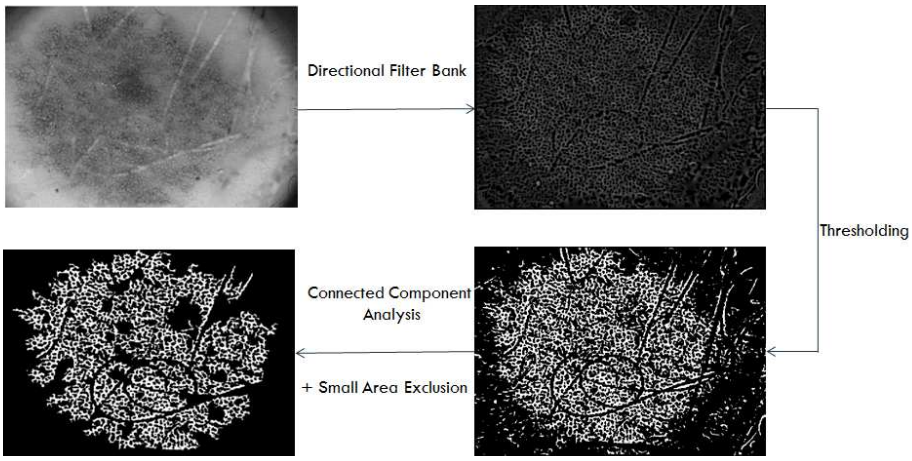


Figure 3.6: Pigment network detection process.

Figure 3.6 illustrates the steps involved in the detection of the the pigment network.

3.7 Lesion Classification

The third block of the proposed detection system aims to assign a binary label to a given image, stating whether or not pigment network is present in the image. Two of the main properties of pigment network are its density and regular spatial distribution. These two properties can be used to identify the images where it is present.

In an attempt to characterize the previous properties, five different features were defined:

(i) Network/Lesion ratio: This value compares the area of the network R with the area of the whole lesion

$$Network/Lesion = \frac{A(R)}{A(L)}, \tag{3.9}$$

where $A(R)$ is the area of the detected network R and $A(L)$ is the area of the segmented lesion L. This ratio requires that each lesion is segmented. Therefore, each of the dermoscopy images was manually segmented by an experienced dermatologist (see Figure 3.7 for an illustrative example).

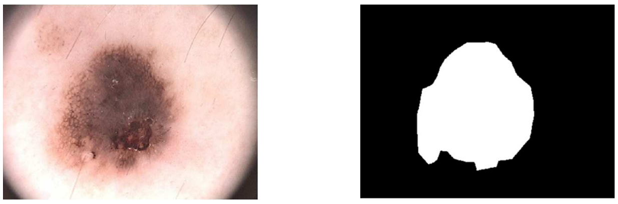


Figure 3.7: Lesion segmentation: original image (left) ; binary segmentation mask (right).

(ii) Network/Regions ratio: This ratio compares the total area occupied by network R with the total area of the pigment network regions, herein identified as B. The latter correspond to the area occupied by the network and its holes, and is obtained by applying a simple morphological filling process to each of the network regions R_c. This lead to a set of regions B_c, as illustrated on Figure

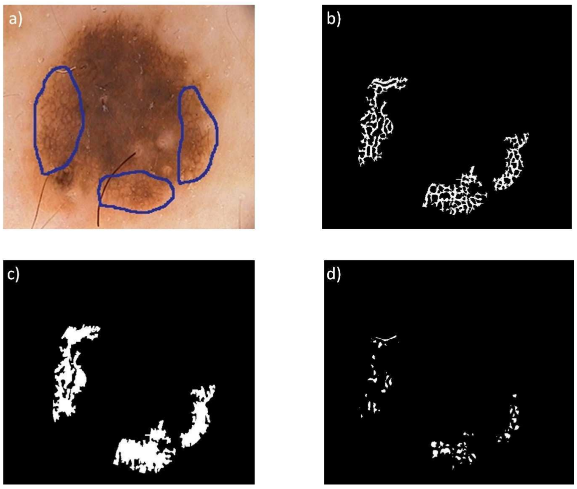


Figure 3.8: Pigment network masks: a) original image with pigment network regions highlighted; b) *net-mask*, the output of the detection block; c) *pigment network regions-mask* and d) *holes-mask*. Images from [16]

3.8(c)). A *pigment network regions-mask* B is defined as the union of all the regions

$$B = \bigcup_{c} B_c, \tag{3.10}$$

The Network/Regions ratio is then obtained as follows

$$Network/Regions = \frac{A(R)}{A(B)}, \tag{3.11}$$

where $A(R)$ is the area of the detected network R and $A(B)$ is the area of all the pigment network regions present in B.

(iii) Number of holes: This feature is the total number of holes in the detected mesh R. A *holes-mask* H can be easily obtained by subtracting R from B (see Figure 3.8(d)).

(iv) Holes/Lesion ratio: This feature is the ratio between the number of holes and the area of the lesion

$$Holes/Lesion = \frac{\#H}{A(L)}, \tag{3.12}$$

where $\#H$ is the number of holes in the *holes-mask* H and $A(L)$ is the area of the segmented lesion L.

(v) Holes/Region ratio: This feature is the ratio between the number of holes and the the area of the pigment network regions

$$Holes/Region = \frac{\#H}{A(B)}, \tag{3.13}$$

where $\#H$ is the number of holes in the *holes-mask* H and $A(B)$ is the area of all the pigment network regions present in B.

These five features are organized in a feature vector and used to train a classifier to detect the presence of pigment network in skin lesions. The AdaBoost classifier [72] is employed for this purpose due to its ability to select a subset of the most appropriate features.

Table 3.1 shows the mean values of the individual features for both classes. These values demonstrate that the features used are appropriate descriptors, since most of them change between classes.

Figure 3.9 shows the histograms for three different features. It is also possible to see that the two classes have different distributions for the three represented features.

Table 3.1: Mean values of the individual features for both classes.

Features	i	ii	iii	iv	v
with pigment network	0.218	0.486	175.3	0.0015	0.003
without pigment network	0.068	0.371	32.7	0.0004	0.002

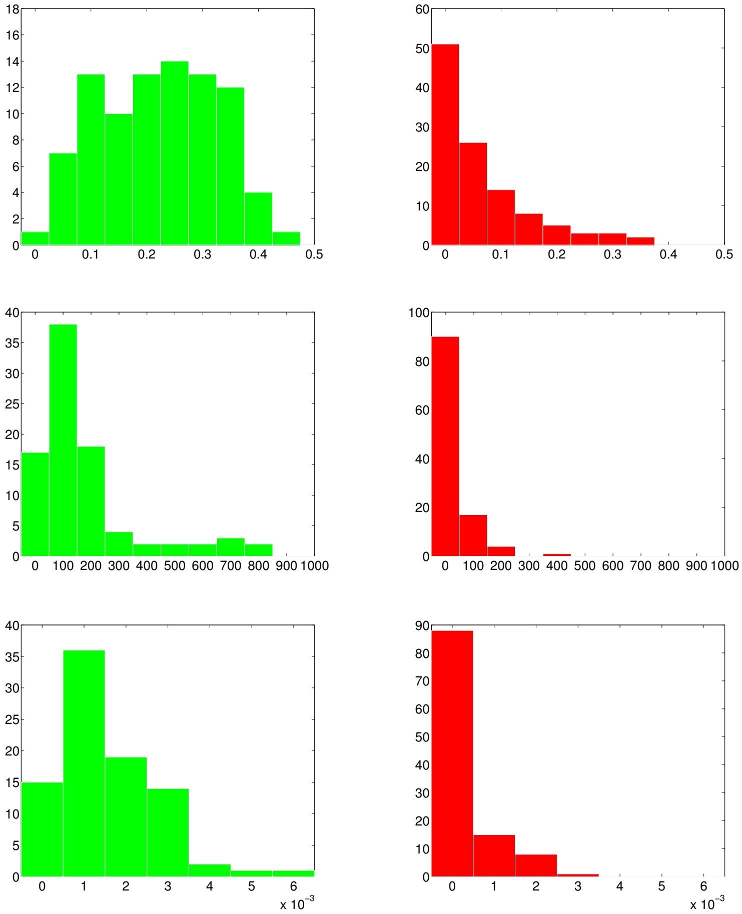


Figure 3.9: Histogram of features for lesions *with pigment network* (green) and *without pigment network* (red): features i (first row), iii (second row) and iv (last row). Image from [16]

3.8 Experimental Results

3.8.1 Dataset and Evaluation Metrics

The proposed algorithm was tested on a dataset of 200 dermoscopy images (88 with pigment network and 112 without) from the database of hospital Pedro Hispano, Matosinhos. These RGB images are stored in BMP and JPEG formats, and were acquired during clinical exams using a dermatoscope with a magnification of 20×. Their average resolution is 573×765. Most of the images used in the development of this method were included in the PH^2 database.

For training and validation purposes each image was classified by an experienced dermatologist, who provided ground truth information. First, each image was labeled as *with* or *without pigment network* (ground truth label). When pigment network was found, the pigment network regions were manually segmented (ground truth segmentation G_T, exemplified in Figure 3.10).

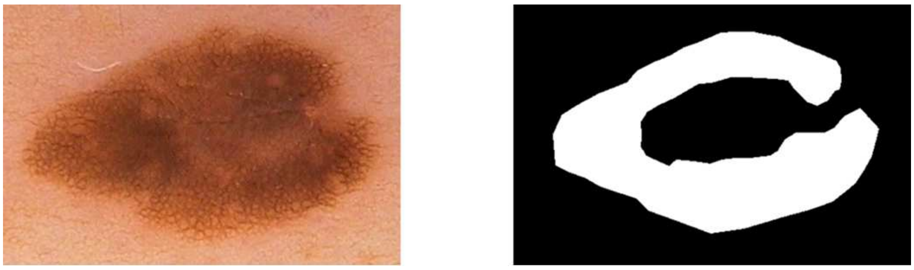


Figure 3.10: Region segmentation: original image with pigment network (left); binary ground truth G_T (right). Image from [16]

Two outputs of the system were evaluated: i) the correct segmentation/identification of pigment network regions; and ii) the binary labeling of lesions as *with* or *without pigment network*. Both were evaluated using appropriate performance measures. The segmentation/identification of pigment network regions must be compared with the ground truth G_T to assess its performance. Two strategies were used to perform the evaluation of the detected regions:

(i) Pixel detection statistics: The binary mask of detected regions B is compared with the ground truth image G_T pixel by pixel. Each pixel is classified as a true positive (TP), true negative (TN), false positive (FP) or false negative (FN). These values are then used to compute the sensitivity (SE) and specificity (SP) as follows

$$SE = \frac{\#TP}{\#TP + \#FN} \tag{3.14a}$$

$$SP = \frac{\#TN}{\#TN + \#FP}. \tag{3.14b}$$

(ii) Region detection statistics: This method works at the region level, *i.e.*, each region in B is separately compared with the binary ground truth G_T and classified in one of the following classes: correct detection (CD) if the detection region matches (overlaps) one or more regions in G_T, false alarm (FA) if the detected region has no correspondence or detection failure (DF) if one region in G_T has no correspondence. This procedure is similar to the one proposed by Nascimento and Marques [134]. The three classes are obtained by computing a matrix C, which defines the correspondence between the active regions of a pair of images. Assuming that G_T consists of P regions and that B contains M detected regions, C will be a P×M matrix computed as follows

$$C(c,p) = \begin{cases} 1 & \text{if } A(B_c \cap G_{T_p}) \neq 0 \\ & \qquad \forall\, p \in \{1,\ldots,\mathrm{P}\}, c \in \{1,\ldots,\mathrm{M}\} \\ 0 & otherwise, \end{cases} \tag{3.15}$$

where $A(B_c \cap G_{T_p})$ is the area of the region $B_c \cap G_{T_p}$. Two auxiliary vectors, which result from adding the number of ones in each line or column, are also computed

$$L(c) = \sum_{p=1}^{P} C(c,p) \quad c \in 1, \ldots, M, \tag{3.16}$$

$$K(p) = \sum_{c=1}^{M} C(c,p) \quad p \in 1, \ldots, P. \tag{3.17}$$

Each detected region B_c is classified as CD if $K(p) \geq 1$. This rule ensures that detected regions, which are actually the union of two regions of G_T, are only accounted once as CD. Detected regions that result from a split of a ground truth region are independently classified as CD. The number of FA and DF are determined by inspecting the number of empty columns or lines in C respectively, which can be easily achieved using vectors L and K [134].

The performance of the lesion/image classification algorithm (with or without pigment network) is assessed by comparing the output of the classifier trained in the lesion classification step with the lesion ground truth described previously. Each lesion's label can be regarded as a TP, FP, TN or FN and these values are used to compute the SE and SP.

All of the aforementioned statistics were computed using a 10-fold cross validation approach.

3.8.2 Results

The proposed algorithm was applied to the set of dermoscopy images described in the previous section. These images were used to tune all the thresholds and parameters, with a 10-fold cross validation method. In the following subsections some of the chosen parameters will be discussed and results for the performance of detection system will be shown, both for region segmentation and lesion classification.

3.8.2.A Assessment of Region Segmentation

Figure 3.11 displays the receiver operating curves (ROC) for the region detection statistics, using different values of A_{min}, N and T_R. These parameters were selected since they belong to the core block of the proposed detection system. The ROC curves were obtained by varying one of the parameters while maintaining the other two constant and equal to the reference values: $A_{min} = 900\, pixels$, $N = 9$ filters and $T_R = 0.0185$.

These graphics suggest a high dependency of the DF vs. FA ratio with both the A_{min} and T_R values. The number of directional filters N has a smaller impact in this ratio, when compared with the other two, provided that $N \geq 9$. It is also shown that the reference values for N, T_R, and $A_{min} = 900$ are the ones that lead to the best results. Table 3.2 summarizes the number of CD, FA, and DF for the region detection problem. A pixel detection statistics was also performed over the detected regions. Table 3.2 displays the results obtained. The low value obtained for SE can be explained by the analysis of Figure 3.12: the manually segmented regions provided by an expert are wider than the detected ones. Therefore, it can be assumed that it is not exclusively related to the existence of DF.

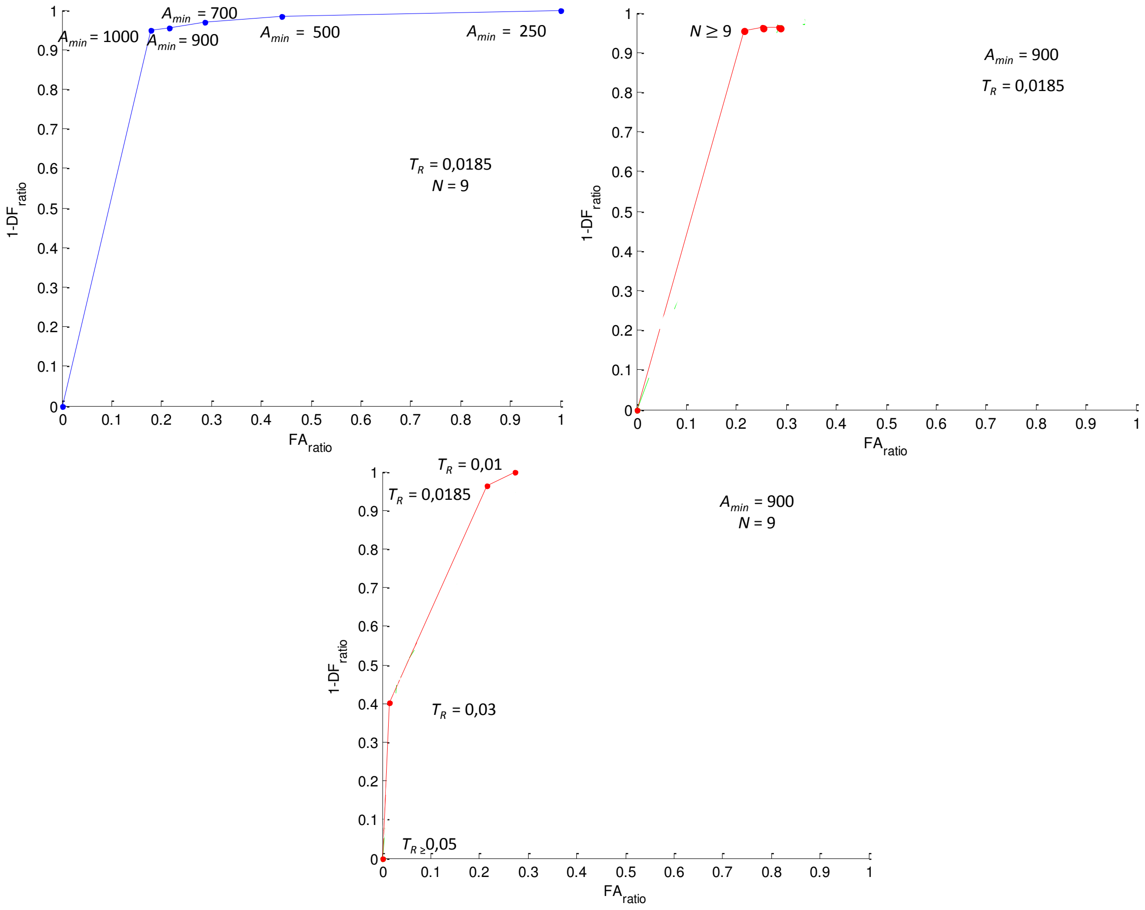


Figure 3.11: ROC curves for (from left to right):$A_{min} \in \{500, 700, 900, 1000\}$, maintaining $N = 9$ and $T_R = 0.0185$; $N \in \{9, 18, 36\}$, maintaining $T_R = 0.0185$ and $A_{min} = 900$; $T_R \in \{0.01, 0.0185, 0.03, 0.05, 0.1\}$, maintaining $N = 9$ and $A_{min} = 900$. Figure adapted from [16].

This evaluation method is less informative than the previous one, where the performance assessment is done at a region level.

Table 3.2: Assessment of pigment network region segmentaion: number of correct detections (CD), false alarms (FA), and detection failures (DF), as well as pixel based statistics for $A_{min} = 900$, $N = 9$ and $T_R = 0.0185$.

Evaluation method	Results
Region statistics	CD = 346
	FA = 358
	DF = 6
Pixel statistics	SE = 57.6%
	SP = 85.4%

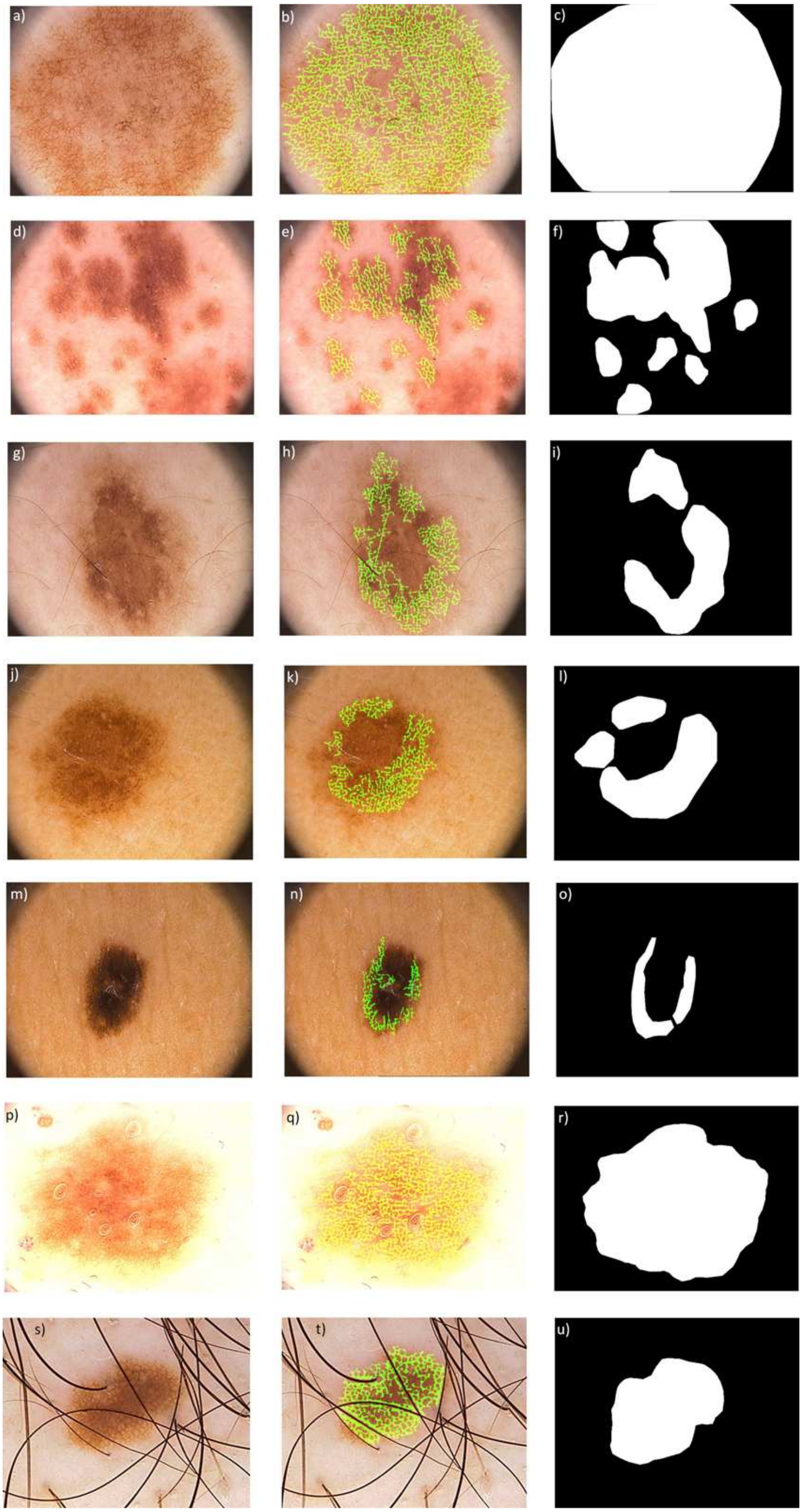


Figure 3.12: Average frequency of selection for individual features.

3.8.2.B Assessment of Lesion Classification

Table 3.3 shows the performance of the lesion classification block using the parameters selected in the previous subsection.

Table 3.3: Statistical results for the lesion classification algorithm for $A_{min} = 900$, $N = 9$, and $T_R = 0.0185$.

Results	#Lesions
SE = 91.1%	88
SP = 82.1%	112
Accuracy = 86.2%	200

Figure 3.12 illustrates the performance of the algorithm for seven images that were deemed as *with pigment network* by an expert. Each one of these images was classified by the developed system as *with pigment network* and the respective detected network is represented in the figure. In order to allow a better comparison with the medical ground truths G_T, these are also displayed. These qualitative results show that the proposed detection system performs well, even in images that can be considered difficult due to the subtlety of the pigment network pattern (see Figures 3.12(g-i) and 3.12(j-l)), the reduced amount of network present (see Figures 3.12(m-o)) or the presence of artifacts (see Figures 3.12(p-r) and 3.12(s-u), bubbles on top of the lesion and hair artifacts).

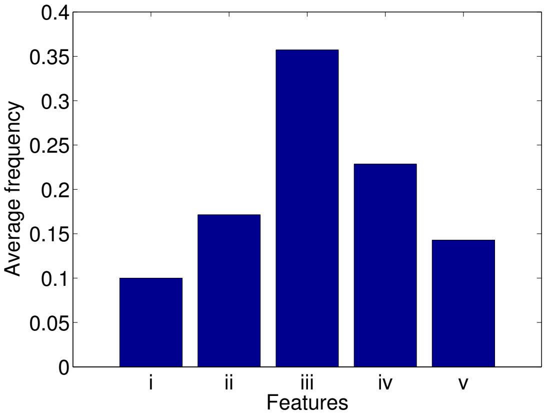


Figure 3.13: Average frequency of selection for individual features.

Figure 3.13 shows a histogram of the average frequency of features selected by the boosting algorithm [72] in the 10 folds. Although all features are selected and seem to play an useful role in the classification task, the most selected feature is ”Number of holes” (feature iii) and the second most selected feature is the ratio Number of holes/Lesion (feature iv), which suggests that these two features are very discriminative. A direct comparison between this work and most of the others found in literature is not possible due to several constraints: i) the outputs of some of the methods are either different from the ones obtained in this paper or not shown; ii) each works uses a different dataset; and iii) the goal of the work is not the same. Nonetheless, it is possible to establish a theoretical comparison with works that have similar goals [6, 159]. The proposed method achieves better lesion classification scores than the ones described in [6]. On the other hand, [159] reports a better accuracy value. Nonetheless, it is important to stress that the methodology described in this work is unique and different from related works, mainly due to its ability to segment the pigment network regions in the imag. This makes it possible for the dermatologist to inspect and validate the output of the system.

3.9 Conclusions

This chapter proposes an algorithm to detect the pigment network in dermoscopy images. The proposed algorithm achieves interesting performances on a dataset of 200 dermoscopy images (88 with pigment network): SE = 91.1% and SP = 82.1%. Besides classifying the lesion as *with* or *without pigment network*, the system is also capable of providing relevant medical information regarding the location of pigment network, since its output is a network mask that highlights the lines of the network. From this mask, it is also possible to extract the final pigment network region as well as the holes of the mesh.

Despite the promising results for network detection, it was not possible to reliably discriminate melanocytic and non-melanocytic lesions. Although the idea of using pigment network to distinguish the two types of lesions is supported by medical findings, the truth is that this structure is not visible in all melanocytic lesions. This means that the decision can not be made using only pigment network as a separating criteria. It seems that expert clinicians use other sources of information when classifying the image as melanocytic or not.

4

A CAD System Based on Global Features

Contents

4.1 Motivation

The previous chapter describes an algorithm for the detection of pigment network, which is a contribution for the distinction between melanocytic and non-melanocytic lesions. This chapter will focus on the problem of melanoma detection, with the assumption that the lesion is melanocytic.

CAD systems for melanoma diagnosis share a similar structure: i) lesion segmentation; ii) feature extraction; iii) (optional) feature selection; and iii/iv) lesion classification [103]. During the feature extraction step it is common practice to describe the lesion as whole, *i.e.*, a single feature vector is computed to characterize the entire lesion (global features). The extracted features are often inspired by the medical ABCD rule [182], and can be divided into four categories: shape, color, texture, and symmetry. Different descriptors have been proposed to represent each of the aforementioned categories. However, after a thorough search of the literature, two things remain unclear: i) whether the four types of features are equally relevant; and ii) what is the best descriptor for each feature category. Moreover, most of times the methodology used to compute the descriptors is poorly described, which makes it difficult to reproduce the experiments. Finally, the systems are developed using different datasets, which makes it impossible to perform reliable comparisons.

The main goal of this chapter is to evaluate the role of the four types of features, as well as to compare different descriptors. This will provide insightful information about the relative importance of each feature as well as which are the most suitable descriptors. The methodology used to compute each of the descriptors is provided and all the experiments are performed using the publicly available PH^2 dataset [126]. This means that all of the methods and results can be reproduced by other research groups.

4.2 System Overview

Figure 4.1 shows the block diagram of the proposed CAD system.

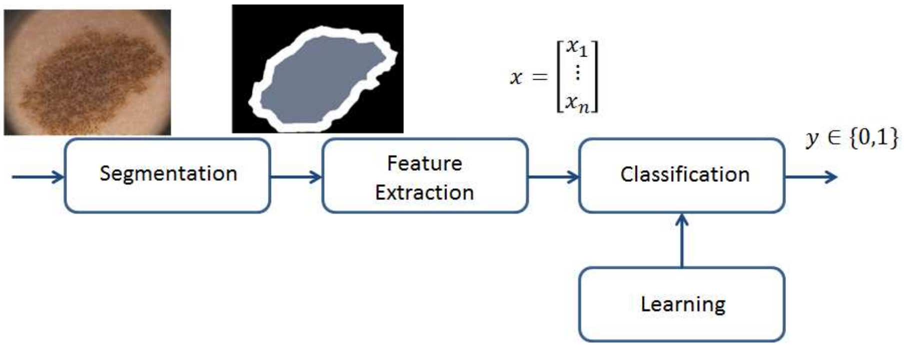


Figure 4.1: Block diagram of a melanoma detection system that uses global features.

Lesion segmentation is performed in order to separate the lesion from the healthy skin. This is a critical step since features will be extracted over the output of the segmentation. In order to prevent melanoma detection errors caused by incorrect segmentations, all the lesions have been manually segmented by an experienced dermatologist. This will also make it easier to interpret the performance of the system for each type of feature.

Since the goal of this work is to understand the role of each type of feature (*e.g.*, color and symmetry) and assess which are the best descriptors, a single type of feature is extracted in the *feature extraction* block. The studied features belong to the four classes presented before: symmetry, shape, color, and texture. The descriptors analyzed for each type of features will be defined in the following section.

Dermatologists pay a special attention to the border of the lesions, *e.g.*, if there are abrupt color transitions or irregular structures. In order to include this knowledge in the CAD system, color and texture features are separately extracted from two regions of the lesion, namely the border and the inner part. The splitting of the lesion into these two regions is performed by eroding the lesion segmentation mask with a disk of radius r, which was experimentally set to be $\frac{1}{10}$ of the lesion's smallest axis. Figure 4.1 shows the splitting of the lesion in the two regions. It was experimentally found in [19] that dividing the lesion into border and inner parts leads to a slight improvement of classification results using color and texture features. Shape and symmetry features are computed using the whole lesion.

The learning phase is performed using the PH^2 dataset. Four classifiers are compared in this work: AdaBoost [72], support vector machines (SVM) [47] with a radial basis function (RBF) kernel, k-nearest neighbor (kNN) [60], and random forests [29]. These algorithms have been selected since they are some of the most popular classifiers and have been used in previous works related with melanoma detection [103, 139].

It is assumed that the described system will only be applied to melanocytic lesions.

4.3 Image Features

This section defines the features and corresponding descriptors that are investigated. As stated in Section 2.3.1, it is common to select categories of features that can be associated with the ABCD rule of dermoscopy [182]. Therefore, four different types of features will be used to describe skin lesions: symmetry, shape, color, and texture. Moreover, several representative descriptors are considered for each type of feature, and the best descriptor is selected. Some of the descriptors investigated in this thesis have been used for the first time in this context, and will be pointed out in the following sections.

4.3.1 Color Features

In the ABCD rule, color analysis is performed by counting how many, out of six clinically relevant colors, can be found in a lesion [182]. Inspired by this criteria, different approaches have been adopted to automatically describe the colors of a given lesion [118]. One of the most popular strategies is to use statistical descriptors such as the mean, standard deviation, min/max values, and entropy computed for each color channel [121]. Other approaches try to characterize the color distribution using quantization techniques. Examples of these methods are the use of color histograms and clustering strategies [39, 103].

Each of the aforementioned descriptors can be computed in one or more color spaces. However, the choice of the most suitable color space is still an open question, and several works deal with this problem by computing color features in more than one space. Dermoscopy images usually come in the RGB format, which means that each pixel is represented by a mixture of three colors: red (R), green (G), and blue (B). RGB has a series of drawbacks: i) it is not perceptually uniform; ii) it is not device independent; and iii) exhibits a high correlation among the three color channels [184]. To overcome these issues, other color representations are usually used. A common strategy is to selected color spaces that have a relation with the human perception of color or that are biologically inspired. An example of such spaces is HSV, which belongs to the family of phenomenal color spaces. These spaces characterize colors in a way similar to that of the human mind, namely each color is characterized by a (H)ue, (S)aturation, and (V)alue [99, 184]. The downside of phenomenal color spaces is that they still lack perceptual uniformity. To tackle this issue one can use CIE color spaces, where the description of color is strongly correlated with the human visual perception [99, 184]. CIE L*a*b* was selected to be used in this work. Finally, an alternative to the previous spaces is the opponent color space. This is a biologically inspired color space, where each of the channels represents respectively the red-green differences, the blue-yellow differences, and the luminance [28, 187]. It is possible to obtain each of the three alternative channels from the traditional RGB, using appropriate transformations [184, 187]. In this work, the RGB, HSV, L*a*b*, and opponent color spaces are tested in order to determine which is the most suitable. The opponent color space has been used for the first time in dermocopy image analysis in this thesis, as reported in [19].

The most popular color descriptors in dermoscopy image analysis are *1-D color histograms*, due to their ability to describe the color distribution and variability inside a lesion. These descriptors have also been used in all sorts of problems related with image classification and object recognition. The idea in this thesis is to compute a set of three histograms for the border and inner parts of the lesion (recall Section 4.2). Each histogram corresponds to one of the color channels and has M_c bins. The color histogram associated to the color channel I_c, $c \in \{1,2,3\}$ is given by

$$h_c(i) = \frac{1}{N} \sum_{x,y} b_c(I_c(x,y)) \quad i = 1, ..., M_c, \tag{4.1}$$

where N is the number of pixels that belong to the inner or border part of the lesion, i is the histogram bin, and $b_c(i)$ is the characteristic function of the $i-th$ bin. This function is equal to one only when $I_c(x,y)$ belongs to the $i-th$ bin and zero otherwise. After computing the histograms for the border and inner part, these values are all concatenated into a single feature vector.

4.3.2 Texture Features

Texture features characterize the spatial distribution of intensity in a given image. Thus, this type of features has been used to loosely represent the dermoscopic structures, as it would be performed in the D step of the ABCD rule. Different texture descriptors have been applied to the melanoma detection problem [103, 121]. These can range from intensity distribution descriptors (*e.g.*, entropy) to filtering techniques (such as Gabor filters [10]) or Haralick descriptors (co-ocurrence matrix [84]).

This thesis explores filtering approaches, namely Gabor filters [10] and gradient-related features [50] (amplitude and orientation histograms). The latter have been applied for the first time to dermoscopy image analysis in [19]. Since texture descriptors are computed over the gray scale image, the first step is to convert the color image into gray scale. This task is accomplished by selecting the RGB color channel with the highest entropy value. Then, the descriptors are computed as follows

- **Gabor filters:** These filters have been widely used for texture classification [10] and edge detection [81]. Furthermore, they have also been used in the melanoma detection context [139]. The impulse response of the $i-th$ ($i = 1, ..., N$) Gabor filter of a certain filter bank of size N is given by [81]

$$h_i(x,y) = e^{\frac{x'^2+\gamma^2 y'^2}{2\sigma_G^2}} \cos\left(2\pi\frac{x'}{\lambda}+\phi\right) \tag{4.2}$$

  where γ is an aspect ratio constant, σ_G is the standard deviation, λ is the wavelength, ϕ is the phase of the filter and x', y' are obtained from rotating (x, y) as follows.

$$x' = x\cos\theta_i + y\sin\theta_i\,, \tag{4.3a}$$

$$y' = y\cos\theta_i - x\sin\theta_i\,, \tag{4.3b}$$

  The angle amplitude $\theta_i \in [0, \pi]$ determines the orientation of the filter i and the step between two consecutive orientations is $\frac{\pi}{N}$.

  Gabor filters are applied as follows. First, the gray scale image $I(x, y)$ is convolved with each of the $h_i,\ i = 1, ..., N$ filters in the bank

$$J_i(x,y) = h_i(x,y) * I(x,y)\,. \tag{4.4}$$

  Then, an energy measure is computed for each of the outputs $J_i(x, y)$

$$E_i = \sum_x \sum_y J_i(x,y)^2\,. \tag{4.5}$$

  The image is now characterized by a feature vector $[E_1, E2, ..., E_N]^T$.

- **Amplitude histogram:** The first step is to compute the gradient $g(x, y) = [g_1(x, y)\ g_2(x, y)]^T$ of the image. This is performed using the Sobel masks. Then, the amplitude of the gradient is computed as follows

$$\| g(x,y) \| = \sqrt{g_1(x,y)^2 + g_2(x,y)^2}. \tag{4.6}$$

  Finally, a histogram with M_A bins is determined, using an expression similar to (4.1) with I replaced by $\| g(x, y) \|$.

- **Orientation histogram:** The orientation of the gradient is computed as follows

$$\varphi(x,y) = \tan^{-1}\left(\frac{g_2(x,y)}{g_1(x,y)}\right)\,. \tag{4.7}$$

  As in the case of the amplitude, the histogram with M_φ bins is computed using (4.1).

4.3.3 Shape Features

Dermatologists consider that most benign lesions have small dimensions and a shape similar to a circle [9]. Thus, the goal of shape features is to characterize these two aspects of the lesions. Measures such as the area, perimeter, and compactness index have been widely incorporated in CAD systems [103, 121]. Moreover, shape features also provide information about the border of the lesion (recall the importance of border irregularities in the ABCD rule (Table 2.2). Some examples of the descriptors that provide this kind of information are the convex hull and bounding box descriptors, fractal dimension, and the irregularity index [103].

In this work, three sets of shape descriptors are investigated:

- **Simple shape descriptors:** These descriptors comprise five simple shape features, namely, the area (A), compactness index, minor and major axis length, and rectangularity. The area corresponds to the number of active pixels of the binary segmentation mask. The compactness index
$$C = \frac{4\pi A}{P^2}, \tag{4.8}$$
characterizes the similarity between a lesion shape and a circle with the same perimeter (P). The minor and major axis lengths are obtained by approximating the lesion region by an ellipse, using principal component analysis. Finally, rectangularity, is the ratio between the area of the lesion and the area of the smallest rectangle that contains the lesion.

- **Hu's invariant moments [87]:** These moments are a set of seven descriptors that are invariant to rotation, translation, and scaling, having been applied to object recognition tasks from the object silhouette. They are obtained through a non-linear transformation of the normalized central moments or order p, q [92]
$$\mu_{pq} = \sum_{x,y} (x - \overline{x})^p (y - \overline{y})^q B(x, y), \tag{4.9}$$
where $(\overline{x}, \overline{y})$ are the coordinates of the lesion's center and $B(x, y)$ is the binary segmentation mask. Hu's moments can be related with different geometrical properties, such as the moment of inertia [87].

- **Wavelet invariant moments [169]:** These moments are also invariant to scaling, rotation, and translation. Invariance to scale and translation is obtained by performing a change of variables and the conversion of the binary mask $B(x, y)$ to polar coordinates. After computing the moments, these are made rotation invariant through the computation of their norm.

4.3.4 Symmetry Features

The symmetry of a skin lesion is an important cue to asses its degree of malignancy [9]. Symmetric lesions with respect to two axis tend to be benign, while highly asymmetric lesions, produced by an abnormal increase of melanocytic cells, are often melanomas [1]. The clinical assessment of the symmetry is performed regarding three criteria: shape, color, and dermoscopic structures [182].

Most of the CAD systems try to characterize the symmetry of the lesion regarding the first criteria. The analysis of shape symmetry is performed by dividing the segmentation mask into an even number of slices and then comparing the area differences between overlapping ones [38]. Symmetry regarding color and dermoscopic structures is also assessed, but the amount of works that focus on these two aspects is much smaller [163, 166]. These works start by diving the lesion into small patches, which are then characterized using either color or texture features. Finally, differences between the features of symmetric patches are assessed and used to describe the degree of symmetry of the lesion.

In this thesis, the three types of symmetry are compared. To assess shape symmetry the segmentation mask of the lesion is divided into an even number of slices, with a common vertex at the lesion centroid. Then, the opposite slices are flipped, overlapped, and the degree of symmetry is computed as follows

$$s = \frac{2A_{ij}}{A_i + A_j}, \tag{4.10}$$

where A_{ij} corresponds to the area of intersection between the two slices and A_i, A_j are the areas of the opposite slices i and j. This procedure is exemplified in Figure 4.2, where the lesion was divided into 16 slices and A$_{ij}$ corresponds to the red region and A$_i$, A$_j$ correspond to the blue and yellow regions.

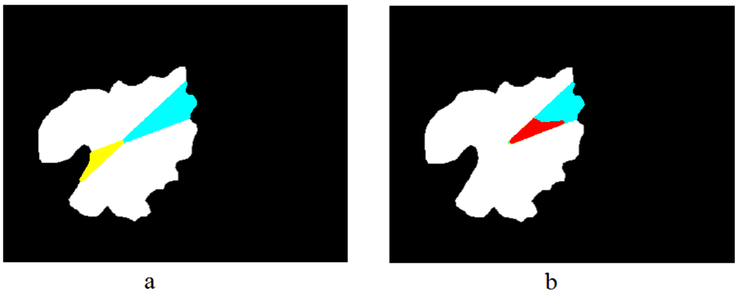


Figure 4.2: Analysis of the symmetry between opposite slices: (a) two opposite slices of the lesion and (b) overlap (red) between one of the slices (blue) and the mirrored image of its opposite slice (yellow). Images from [152].

The strategy used to assess color and texture symmetry has been reported for the first time in [152]. First, the two principal axis of the lesion are found by applying principal component analysis (PCA) [96] to the binary segmentation mask. Then, a regular grid oriented according to the two principal axis is placed on top of the lesion, as exemplified in Figure 4.3. The nodes of the grid are computed as follows.

$$\mathbf{x}_{jk} = j\Delta_1\mathbf{v}_1 + k\Delta_2\mathbf{v}_2 + \bar{\mathbf{x}}, \qquad j, k \in \mathbb{Z}, , \tag{4.11}$$

where $\bar{\mathbf{x}}$ is the centroid of the lesion, λ_i and $\mathbf{v}_i$ are respectively the eigenvalues and eigenvectors obtained using PCA, and Δ_i is set to be

$$\Delta_i = c\sqrt{\lambda_i}, \qquad i \in \{1, 2\}. \tag{4.12}$$

c is a constant used to define the distance between the grid nodes.

The following step consists of extracting a set of features to characterize each of the grid blocks. Each block is processed only if at least 25% of its area corresponds to lesion. Depending on the

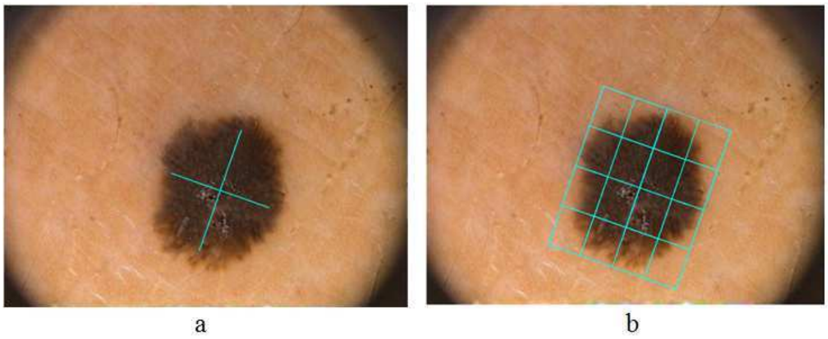


Figure 4.3: Regular grid of nodes: (a) principal axis; (b) grid nodes. Images from [152].

type of symmetry, color or texture features are extracted. The color features used are the *mean color vectors*, computed for each of the four color spaces (RGB, HSV, L*a*b*, and opponent). The texture features are the ones described in Section 4.3.2.

After describing each block by a feature vector, it is necessary to compare them and assess the degree of asymmetry of the lesion. This task is accomplished by first defining the sets S_k, which comprise all of the pairs of symmetric blocks (i, j) according to the $k = 1, 2$ symmetry axis *i.e.*, the principal axes. Then, a distance measure is applied to the features of symmetric blocks, starting with the symmetry axis $\mathbf{v}_1$. The metric used is the Euclidean distance

$$d(\mathbf{x}_i, \mathbf{x}_j) = \| \mathbf{x}_i - \mathbf{x}_j \|, \tag{4.13}$$

where $\mathbf{x}_i, \mathbf{x}_j$ are feature vectors associated with the pair (i, j) of symmetric blocks. Finally, statistical measures are computed for each of the sets, namely: mean μ_k, standard deviation σ_k, maximum M_k, and minimum m_k. Each of these measures is computed as follows.

$$\mu_k = \frac{1}{\#S_k} \sum_{(i,j)\in S_k} d(\mathbf{x}_i, \mathbf{x}_j), \tag{4.14}$$

$$\sigma_k^2 = \frac{1}{\#S_k} \sum_{(i,j)\in S_k} (d(\mathbf{x}_i, \mathbf{x}_j) - \mu_k)^2, \tag{4.15}$$

$$M_k = \max\{d(\mathbf{x}_i, \mathbf{x}_j) : (i, j) \in S_k\}. \tag{4.16}$$

$$m_k = \min\{d(\mathbf{x}_i, \mathbf{x}_j) : (i, j) \in S_k\}, \tag{4.17}$$

The final color/texture symmetry descriptor is obtained by concatenating all of the previous statistics: $[\mu_1, \sigma_1, M_1, m_1, \mu_2, \sigma_2, M_2, m_2]$.

4.4 Classification Algorithms

Different classification algorithms were used in the development of CAD systems for melanoma diagnosis, as reported in Table 2.4 [103, 121]. However, it is not clear if one method outperforms the others. Since each classification algorithm has strong and weak points, it is not easy to select only one to perform all of the feature and descriptor comparisons. Therefore, four different algorithms were applied in this chapter.

- **k-Nearest Neighbor (kNN) [60]** - This algorithm has a very simple formulation: given a training set of patterns (feature vectors) for which the classes are known, each new pattern will be classified in the same class as that of the closest training pattern (called nearest neighbor). The comparison between patterns is performed by computing the distance between feature vectors. It is possible to determine the class of the test pattern by taking into account not only one but the k closest training patterns (neighbors). In this case, the class of the pattern will be the one that is more common among the selected training patterns.

  The reason for selecting kNN is two-folded: i) it is simple and easy to implement; and ii) it achieves competitive performances when the size of the dataset and/or feature space is small, as is the case of this work. Two parameters must be optimized for this algorithm: the number of neighbors k and the distance metric used to compare the feature vectors of training and test images.

- **Support vector machines (SVM) [47]** - The goal of SVM is to learn an hyperplane that separates the training patterns of two classes. Since this problem may have multiple (infinite) solutions, another restriction is added: the separation hyperplane must be the one that has the largest distance to the nearest training pattern of each class. This distance is called margin, hence one can formulate the learning problem of SVM as that of finding the optimal hyperplane that maximizes the margin to the training data. Sometimes it is not possible to separate the training data using an hyperplane, since the two clouds of training features overlap. To deal with this difficulty, it is necessary to relax the margin constraint and use a soft-margin formulation instead. In this case, a penalty term with hyperparameter C is added to the optimization problem. This penalty represents the trade-off between increasing the margin size and ensuring that a training pattern is correctly classified. C is usually tunned during the training phase.

  The aforementioned formulation assumes that the training data are linearly separable, which is not the case in most classification tasks. The strategy used to solve this problem is to map the patterns into a high dimension space, where they are linearly separable.Since the strategy used by SVM to learn the optimal hyperplane relies on the computation of inner product between all pairs of training patterns, mapping them to a higher dimension space poses a computational problem. The strategy used to solve this issue is called kernel trick, where a kernel function is used to compute the inner product between the feature vectors in a high dimension, without actually having to map them. Different kernel functions can be found in literature, and some of them require the tunning of one of more parameters. The most popular kernel function is the Gaussian radial basis function (RBF), defined as follows

$$Kernel(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma||\mathbf{x}_i - \mathbf{x}_j||^2), \tag{4.18}$$

where $\mathbf{x}_i, \mathbf{x}_j$ are feature vectors associated with patterns i and j, and γ is a parameter that must be optimized and define the width of the kernel. SVM-RBF has been applied to dermoscopy images with success (*e.g.,* [38]). Therefore, it will also be used in this thesis.

- **AdaBoost [72]** - This algorithm belongs to the family of boosting methods, where the idea is to combine an ensemble of weak classifiers to obtain a stronger one. The strong classifier is obtained as follows.

  1. Assign weights to each training pattern.
  2. For W iterations:
     i) Learn a stump (weak) classifier for each feature component and determine its performance.
     ii) Choose the weak classifier with the lowest classification error.
     iii) Update the weights of the training patterns: increase the weight if a pattern is misclassified and decrease it otherwise.
  3. Define the strong classifier as a linear combination of the W weak classifiers.

  By assigning a higher weight to misclassified training examples, it is possible to increase their importance in the next iteration of the algorithm. Moreover, this kind of formulation ensures that only the most discriminative values of the feature vector are used to build the final classifier.

  AdaBoost has been applied with success to several classification problems, such as face recognition [192], and has also been applied to dermoscopy (*e.g.*, [33]). Its ability to select the best subset of elements from the feature vector as well as requiring the optimization of only one hyperparameter (the number of weak classifiers W), make this algorithm very appealing.

- **Random forests [29]** - Similarly to AdaBoost, random forests are alsobased on an ensemble of classifiers. In this case, an ensemble of T decision trees is learned using a training set of classified patterns. The methodology used to separately learn each tree, according to [29], is the following: i) randomly select a subset of training samples from the original training set; ii) train a decision tree using the random subspace method. The latter means that each split (leaf) of the decision tree will be trained using a randomly selected subset of values from the training feature vectors. After training the T trees, the classification of new data is performed by taking the majority vote over all the trees. During the training phase it is necessary to optimize the value of T.

  Decision trees have been widely used in dermoscopy image analysis (see Table 2.4). However, fewer works explore random forests [75], which have been proposed to deal the tendency of decision trees to overfit training set. This motivated the application of random forests in this thesis.

4.5 Experimental Results

4.5.1 Dataset and Evaluation Metrics

The CAD systems built using the different types of features and descriptors were evaluated using the PH^2 database [126], acquired and annotated at hospital Pedro Hispano, Matosinhos. This dataset contains 200 dermoscopy images, 40 of which are melanomas. These images were acquired during regular clinical practice using a digital dermatoscope with a magnification of 20×. Their format is RGB and the average size is 573×765. All the images were used to evaluate the performance of the different color and texture descriptors. However, shape and symmetry analysis can only be performed in lesions that are almost fully contained in the image. Thus, lesions that intersected the borders of the image by more than 20% were excluded. This lead to the creation of a reduced dataset of 165 images, 12 melanomas (unfortunately, most melanomas were excluded). Tests with color and texture descriptors were performed in both sets while the shape and symmetry tests were carried out using the reduced set only. All the lesions were manually segmented by an expert.

The CAD systems were evaluated using the sensitivity (SE) and specificity (SP) statistics. The former corresponds to the percentage of correctly classified melanomas, while the later is the percentage of correctly classified benign lesions. In order to select the best configuration, a cost index was also computed. This index quantifies the trade-off between SE and SP

$$S = \frac{c_{10}(1-SE) + c_{01}(1-SP)}{c_{10} + c_{01}}, \tag{4.19}$$

where c_{10} is the cost of an incorrectly classified melanoma and c_{01} is the cost of an incorrectly classified benign lesion. It is assumed that an incorrect classification of a melanoma is a more serious error, therefore c_{10} is set to 1.5 and c_{01} is set to 1.

All of the previous statistics have been obtained using 10-fold nested cross validation. This means that the data was divided into 10 folds, with approximately the same number of melanomas and benign lesions. From these folds, 9 are kept for training and validation (selection of hyperparameters) and the 10th fold is used for testing. The testing process is repeated ten times with a different fold, while the training-validation processes are performed nine times for each testing fold. Each time a different fold is kept out for validation. Nested-cross validation is shown in Figure 4.4. The main strength of this procedure is that it ensures that the choice of the best hyperparameters is independent of the test set. During the division of the lesions by the 10 folds, it was necessary to pay special attention to the lesions that were going to be excluded during shape and symmetry analysis.

Some of the descriptors and classification algorithms required the tunning of one or more hyperparameters. The specific hyperparameters of the descriptors have been thoroughly investigated in [19]. For the sake of simplicity, the results presented in the next section were obtained using previously assessed hyperparameters for each of the descriptors, which are summarized in Table 4.1. The tested hyperparameters for each classifier are shown in Table 4.2. The best configurations were selected by grid search during the validation phase of the nested-cross validation method. All of the feature

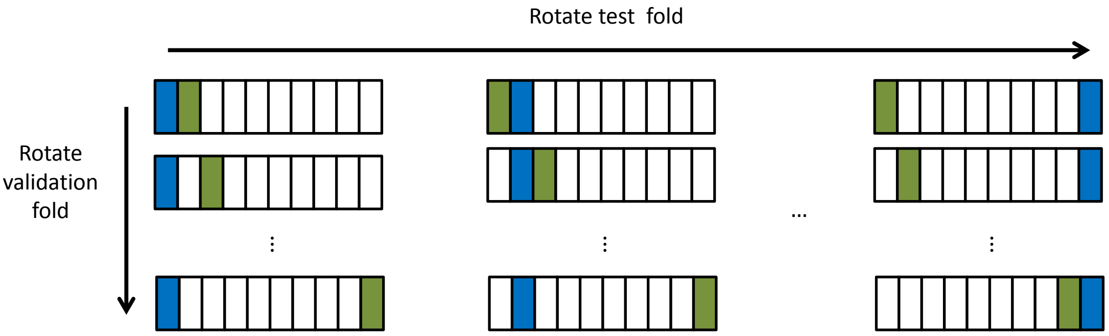


Figure 4.4: Nested cross validation using 10 folds. Each rectangle represents one of the folds, namely training folds (white), validation fold (green), and test fold (blue).

vectors have been normalized as follows.

$$f_i' = \frac{f_i - \mu_i}{\sigma_i}, \tag{4.20}$$

where f_i' is the the normalized i-th component of the feature vector f, f_i is the value before normalization, and μ_i and σ_i are the mean and standard deviation, respectively.

Table 4.1: Descriptors and respective hyperparameter values.

Descriptor	**Parameters**
Color Histograms	Number of bins set to 32.
Amplitude Histogram	Number of bins set to 16.
Orientation Histogram	Number of bins set to 16.
Gabor Filters (4.2)	$\sigma_G \in \{2,4,8\}$, $N = 8$, $\gamma = 0.5$, $\phi = 0\ rad$, and $\frac{x'}{\lambda} = 0.56$.
Shape Symmetry	The lesion is divided into 8 slices.
Color and Texture Symmetry (4.12)	$c = 1.25$.

Table 4.2: Classifiers and hyperparameter values.

Descriptor	**Parameters**
AdaBoost	Number of weak classifiers $W \in \{5, 10, .., 75\}$.
SVM	Soft margin penalty $C \in \{2^{-10}, 2^{-9}, ..., 2^{10}\}$. RBF parameter $\gamma \in \{2^{-16}, 2^{-15}, ..., 2^{16}\}$.
kNN	Number of neighbors $k \in \{3, 5, 7, ..., 35\}$. Comparison distance: Euclidean, Histogram Intersection, Kullback-Leiblller.
Random forests	Number of trees $T \in \{1, 2, 3, .., 50\}$.

It is important to have in mind that the results reported in this chapter were obtained after a significant amount of experiments. For each classifier, the following number of systems was trained

$$\#systems = \#folds_{test} \times \#folds_{validation} \times \#hyperparameters_{conf.} \times \#descriptors. \tag{4.21}$$

This leads to a total of 22950 systems for AdaBoost, 1060290 systems for SVM, 27540 systems for kNN, and 76500 systems for random forests.

4.5.2 Results

Tables 4.3 to 4.6 show the best results obtained for each type of feature using the four different classifiers. For the sake of simplicity, the remaining results can be seen in Appendix A.

Table 4.3: Classification results obtained using the AdaBoost algorithm in the PH^2 and reduced sets. * signals the results obtained with the reduced set. Only the results of the best descritors are shown.

Feature	Descriptor	SE	SP	S
Color	RGB histograms	88%	86%	0.128
	Opponent histograms	90%*	79%*	0.144*
Texture	Amplitude histogram	84%	77%	0.188
	Amplitude histogram	75%*	77%*	0.242*
Shape	Simple shape descriptors	80%*	58%*	0.288*
Symmetry	Texture symmetry	90%*	73%*	0.168*

Table 4.4: Classification results obtained using the SVM algorithm in the PH^2 and reduced sets. * signals the results obtained with the reduced set. Only the results of the best descritors are shown.

Feature	Descriptor	SE	SP	S
Color	L*a*b* histograms	87%	86%	0.134
	HSV Histograms	80%*	80%*	0.200*
Texture	Gabor filters	86%	73%	0.192
	Gabor Filters	75%*	56%*	0.326*
Shape	Simple shape descriptors	85%*	55%*	0.270*
Symmetry	Texture symmetry	70%*	68%*	0.308*

Table 4.5: Classification results obtained using the kNN algorithm in the PH^2 and reduced sets. * signals the results obtained with the reduced set. Only the results of the best descritors are shown.

Feature	Descriptor	SE	SP	S
Color	HSV histograms	89%	85%	0.126
	L*a*b* histograms	100%*	67%*	0.132*
Texture	Amplitude histogram	83%	84%	0.166
	Amplitude histogram	95%*	84%*	0.094*
Shape	Wavelets	65%*	60%*	0.370*
Symmetry	Color symmetry	80%*	75%*	0.220*

These results provide interesting information. First of all, it is important to emphasize the performance of color features when compared with the other types. Color features consistently achieve better classification results, even among different classifiers. The only exception is the case of kNN (see Table 4.5), where the texture descriptor outperforms the color ones for the reduced set. It is interesting to notice that does not seem to exist a preferable color space, since good results are achieved

Table 4.6: Classification results obtained using the random forests algorithm in the PH^2 and reduced sets. * signals the results obtained with the reduced set. Only the results of the best descritors are shown.

Feature	**Descriptor**	SE	SP	S
Color	Opponent histograms	89%	84%	0.130
	HSV histograms	90%*	86%*	0.116*
Texture	Amplitude histogram	84%	80%	0.176
	Amplitude histogram	75%*	73%*	0.258*
Shape	Hu moments	70%*	81%*	0.256*
Symmetry	Color symmetry	90%*	75%*	0.160*

with all of them. Texture features perform well and it seems that the histogram of the gradient amplitude is a good descriptor. This descriptor is the best texture feature in three out of four classifiers. Symmetry features also perform well. Although texture and color symmetry were both selected as the best descriptor, there is not a significant difference in their performances, as can be seen in the remaining results (Appendix A). Overall, shape descriptors seem to be the ones that lead to the worse results. The four classifiers achieve comparable results, but SVM seems to be slightly worse than the other three.

4.6 Conclusions

In this chapter the role of four types global features (color, texture, shape, and symmetry) was investigated using the PH^2 database. This is a necessary task to understand the relative importance of each type of feature and to determine which are the most suitable descriptors. In order to study each descriptor, it was necessary to develop several CAD systems, each one using solely on one type of feature. The final results provided relevant information. First they showed that color features generally outperform the other types of features, while shape features are the ones that lead to the worst results. The different classifiers showed comparable performances in the PH^2 dataset. However, SVM achieved slightly worse results.

The study performed in this chapter is a significant contribution in the field of automatic melanoma diagnosis, since a comparison between the different types of features has never been reported, to the best of our knowledge. This provides insightful information that can be used to develop a CAD system that combines different types of features.

5

The Role of Local Features

Contents

5.1 Motivation

Some of the criteria used by clinicians in the ABCD rule can be efficiently characterized with global descriptors, such as the ones proposed in the previous chapter. Examples are the shape of the lesion and the general color distribution. However, the D criterion, which stands for dermoscopic structures, corresponds to localized texture and/or color patterns. By solely performing a global description of the lesion, one might miss these relevant cues. This chapter tries to overcome this issue using local color and texture features to describe the lesion, and applying the bag-of-features (BoF) model [176] to classify the dermoscopy images. With this model, it will be possible to approximate the automatic analysis to that of a dermatologist, in the sense that the algorithm is also looking for localized relevant aspects.

Local features have been successfully used in several complex image analysis problems, such as scene recognition and object identification (e.g., [95, 108, 176]). However, they have been scarcely explored in the context of melanoma detection. The work developed in this thesis and published in [19] was one of the first to propose the use of local features.

This chapter proposes a CAD system based on local features and considers several color and texture descriptors. Moreover a comparison between global and local features is also performed.

5.2 System Overview - Bag-of-Features model

The main assumption of this chapter is that it is possible to characterize a dermoscopy image using local information, instead of extracting global features. It is assumed that an image is divided into a set of small patches, and that each of these patches is separately represented by a vector of features (color or texture). The goal is to use this local information (patch features) to diagnose melanomas. However, the number of patches extracted among images is not constant, and can be very high. Unfortunately, classic pattern recognition methods cannot cope with this variability. The BoF method provides a way to deal with this problem. According to this model, the local features extracted from the image can be represented by an histogram with a constant number of bins [176]. The main steps of BoF are shown in Figure 5.1. Each of these steps will be addressed in the following sections.

5.2.1 Patch and Feature Extraction

The first step consists of dividing the lesion into small regions. To achieve this goal it is necessary to first find a set of informative keypoints (patch centers) and then extract their corresponding support regions. The identification of keypoints can be either performed using a regular grid (dense sampling) [187], where it is assumed that each keypoint corresponds to one node of the grid, or using one or more keypoint detectors such as Harris Laplace [130] or difference of Gaussians [116] interest points (sparse sampling [187]). Both approaches have been investigated in [18] in the scope of this thesis, and it was concluded that they lead to similar results. In this chapter, the sampling method used is

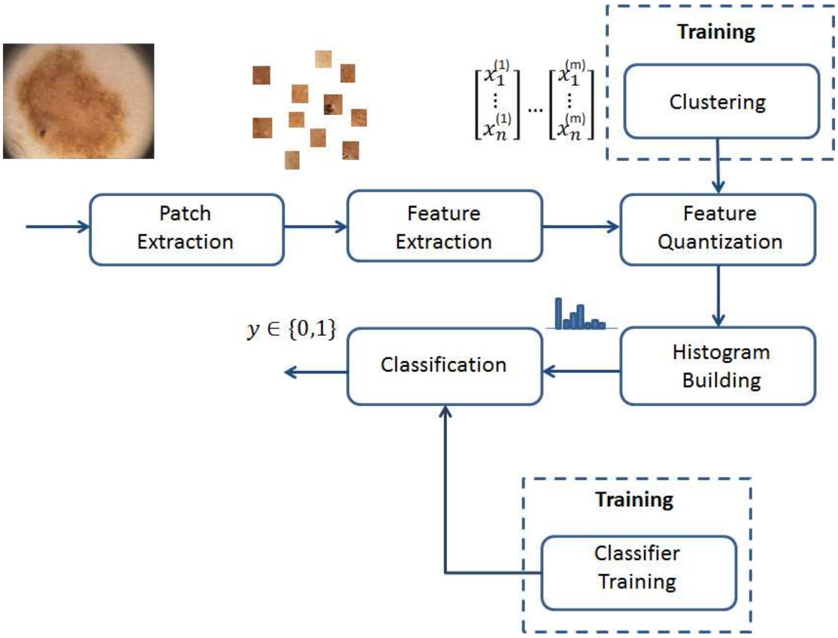


Figure 5.1: Block diagram of a melanoma detection system that uses local features.

dense sampling. Each sampled patch is considered for the posterior steps only if more than 50% of its area corresponds to the lesion. Figure 5.2 shows two examples of the sampling process. The first one is dense sampling and the second is sparse sampling using the Harris Laplace keypoint detector.

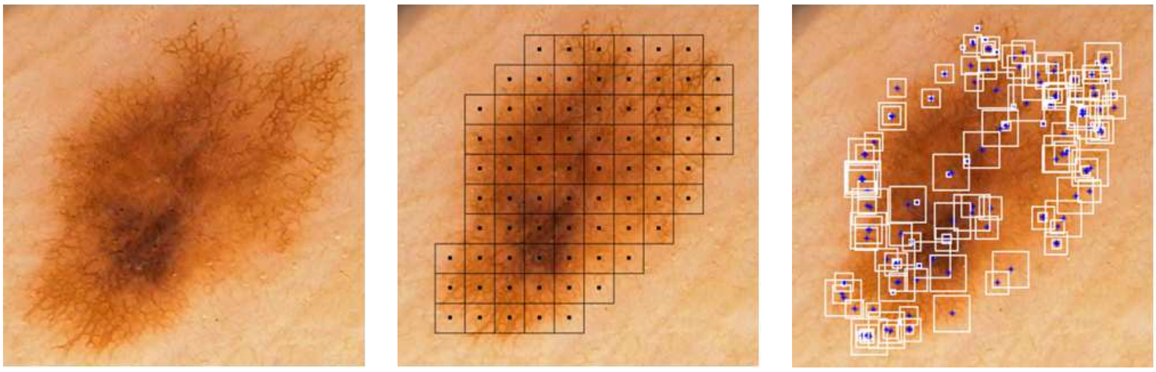


Figure 5.2: Examples of dense sampling: original image (left), dense sampling (mid), and sparse sampling using Harris Laplace (right).

After extracting the patches it is necessary to characterize each of them using a feature vector. In this chapter, the role of color and texture features is investigated, using the same descriptors of Chapter 4. As before, different systems are trained, each solely using one type of descriptor (color or texture). The color descriptors investigated are color histograms computed using the four color spaces: RGB, HSV, L*a*b*, and opponent. In the case of texture, the used descriptors are the amplitude and orientation histograms, and Gabor filters. Other color and texture descriptors have also been studied (*e.g.*, Laws' masks [107], mean color vectors, and SIFT descriptors [116]) and the results are reported in [17, 20]. Figures 5.3 and 5.4 from [20] exemplify the aspect of different color and texture features in different patches of the lesions. The differences among the features of the different patches show that several areas of the lesions possess different color and texture properties. This kind of information would be diluted or even missed, if a global representation was used instead.

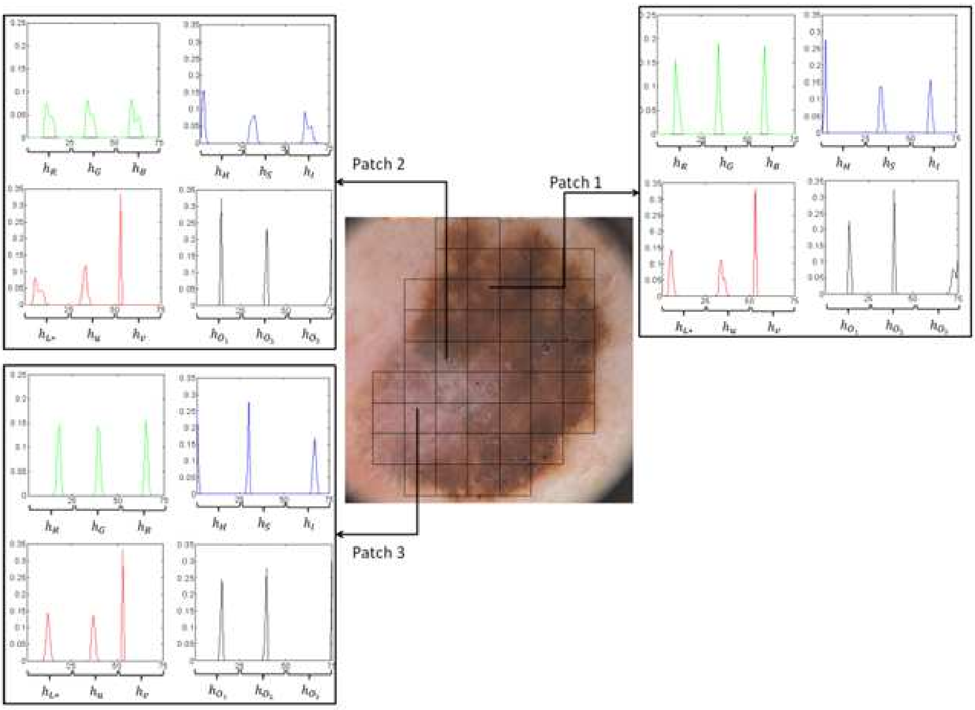


Figure 5.3: Examples of local color histograms using RGB, HSV, L*u*v*, and opponent color spaces. Image from [20]

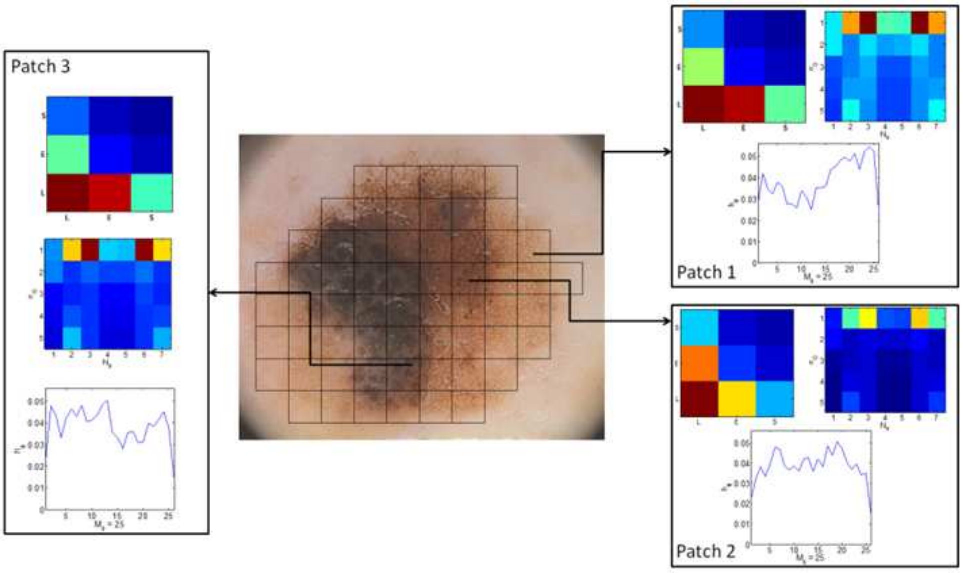


Figure 5.4: Examples of local texture features, namely Laws masks, Gabor filters, and histogram of the gradient orientation. Image from [20]

5.2.2 Clustering

It is not possible to use all the feature vectors to classify the image since the number of features would depend on the image and the dimension of the space would be huge. A simple strategy to solve this problem is to define a dictionary of *visual features* [176], as described in the sequel. The feature vectors of all training images are clustered into a set of groups (typically a few hundred) and a prototype (centroid, often called *visual word*) is extracted from each group. This operation is performed using the k-means algorithm. It is important to stress that while the other blocks occur both in the training and test phases of the model, the clustering step only occurs in the first phase. This is the most demanding operation, in terms of computation time.

5.2.3 Feature Quantization and Histogram Building

Any training or test image will be characterized by an histogram of *visual words* [176]. Each feature vector extracted from the image will be associated to the closest *visual word* using the Euclidean distance. With this information, it is possible to count the number of times each *visual word* occurs in a given image. The results can be represented by means of a histogram that has the same size as the number of prototypes. This histogram is seen as a new feature vector that can be used to characterize the lesion.

5.2.4 Classification

The classification block is different in the training and test phases. During the training phase, the histograms of the training images are used to learn a classification rule, using a supervised classification algorithm. The algorithms used in this chapter are the same ones used in Chapter 4, namely AdaBoost [72], SVM [47] using a RBF kernel, kNN [60], and random forests [29]. During the test phase, the histogram of each new image is classified using the previously learned classification rule. The classifier performance depends on the choice of hyperparameters which are the same as described in Section 4.4.

5.3 Experimental Results

5.3.1 Dataset and Evaluation Metrics

The performance of local features was assessed using the entire PH^2 database (recall Section 2.4). This allows a direct comparison with the results presented in Chapter 5, since they were obtained using the same set of images.

Each CAD system was evaluated using the SE and SP statistics. Furthermore, the cost index (4.19) was also computed in order to select the best CAD configuration for each descriptor. All of the values were obtained using the nested 10-fold cross validation method, described in Chapter 4.

The BoF model, the features and the classification algorithms depend on hyperparameters. Regarding BoF, the specific hyperparameters are the step δ between consecutive nodes of the grid and the size of the dictionary K. The former parameter was set to $\delta = 40$, based on the results obtained in different works [19,20], while the latter was tuned according to the interval $K \in \{100, 200, 300\}$. Regarding the features extracted from each patch, the number of bins used for both color and gradient histograms is 16, and the parameters of the Gabor filters are the same as in Chapter 4 (see Table 4.1). All of the feature vectors were normalized using (4.20). Finally, the parameters of each of the four classification algorithms used were tuned according to the range of values shown in Table 4.2.

In this chapter the number of trained systems is large, as in Chapter 4. The total number of systems per classifier is: 28350 for AdaBoost, 1309770 for SVM, 34020 for kNN, and 94500 for random forests.

5.3.2 Results

Table 5.1 shows the best results achieved by each classifier, using color and texture features. The SE and SP scores of each pair descriptor/classifier can be seen in Appendix B.These results show a couple of things. The first is that local features achieve good scores for all of the classification algorithms, with SVM being the one that performs the worst. Color descriptors outperform texture ones, as in Chapter 4. However, the gap between the two types of features is smaller than when they are used as global features.

Table 5.1: Best classification results obtained using local features.

Classifier	Descriptor	SE	SP	S
AdaBoost	Opponent Histograms	90%	84%	0.124
	Amplitude Histogram	87%	81%	0.154
SVM-RBF	L*a*b* Histograms	87%	79%	0.162
	Orientation Histogram	78%	88%	0.182
kNN	RGB Histograms	100%	76%	0.096
	Gabor Filters	91%	77%	0.146
Random Forests	L*a*b* Histograms	94%	77%	0.128
	Gabor Filters	88%	88%	0.120

Figures 5.5 and 5.6 show the best results obtained using local texture and color features, as well as a comparison with the performance of the same descriptors when used as global features.

According to the results, local features outperform or have a similar performance to the global ones. One of the most relevant results is the improvement in the performances of the orientation histogram and Gabor filters. These descriptors seem more suitable as a local features, since they achieve significantly better scores when used with the BoF model. The gradient amplitude shows a similar performance in both cases, with slightly better results in the case of BoF. Color features achieve good results either as global or local features. Most of the local systems achieve higher SE when compared with their corresponding global systems, without significantly decreasing the SP.

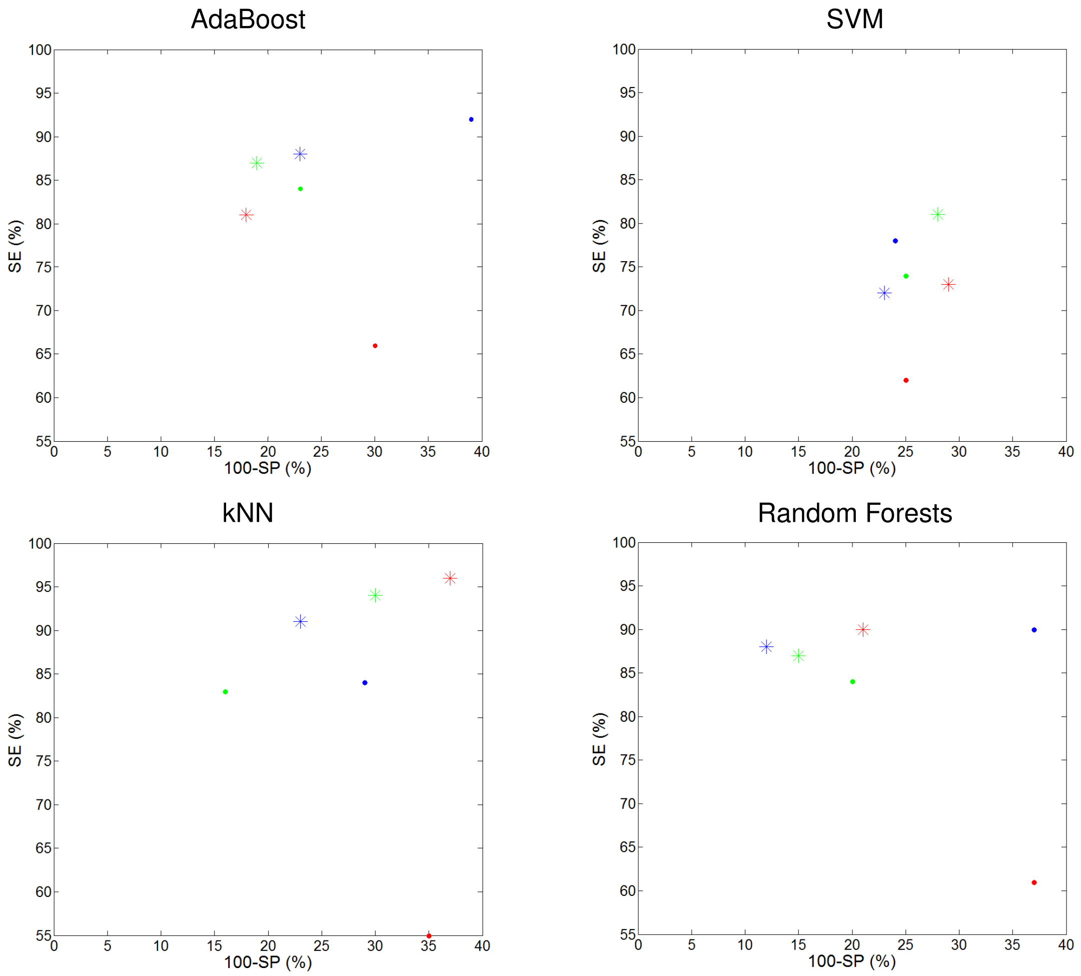


Figure 5.5: Comparison between local ($*$) and global ($\bullet$) texture features. The descriptors are: amplitude histogram (green), orientation histogram (red), and Gabor filters (blue).

5.4 Conclusions

This chapter uses local features to characterize skin lesions and compares the performance of these features against global ones. The results show that local features perform well and some of them significantly outperform global ones. Among all the descriptors, it is important to note the improvement in the performance of texture features.

The obtained results show that, as expected, it is important to characterize certain aspects of the lesions using local features, instead of averaging these cues along the whole lesion. Thus, this kind of information should be included in a final CAD system. Based on the results obtained on this chapter and on chapter 4, the texture of the lesions should be characterized using local features. Color features can be simultaneously used to describe global and local aspects of the lesion, since the results were similar. Nonetheless, it would be preferable to use local color features, in order to effectively characterize localized color structures, such as blue-whitish veil.

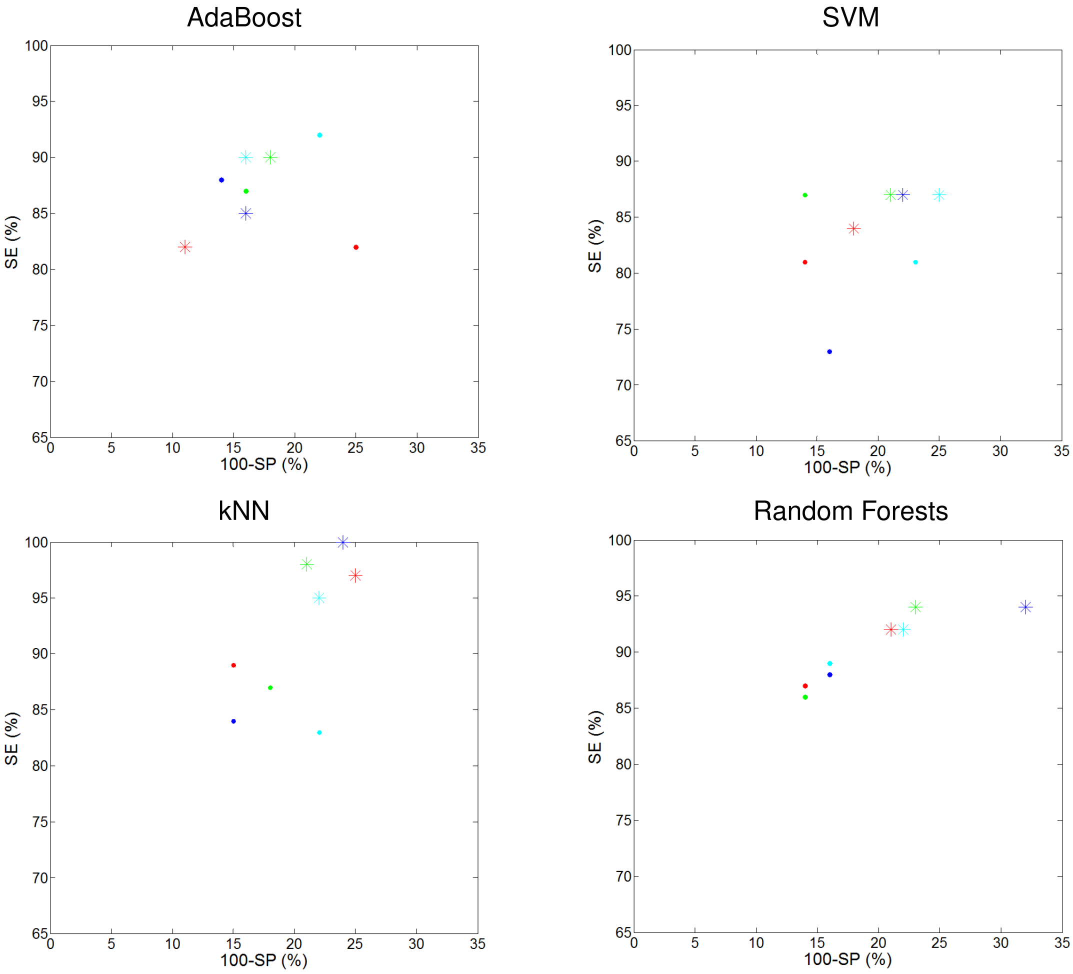


Figure 5.6: Comparison between local (∗) and global (●) color features. The color spaces are: RGB (blue), HSV (red), L*a*b* (green), and Opponent (cyan).

6

Analysis of Images from Multiple Sources/Hospitals

Contents

6.1 Motivation

Most of the pattern recognition CAD systems proposed in literature report accurate results under controlled conditions. However, it is not possible to perform a fair comparison between the different methods, since each system is trained and tested using a different dataset. The datasets are usually obtained using specific acquisition devices and illumination conditions. It is well known that changes in the acquisition setup can alter the colors of an image, as it is exemplified in Figure 6.1. The medical expert has the ability to cope with this variability, but a CAD system will have difficulties, since such variability introduces changes in the values of color-related features (*e.g.*, color histograms). Most CAD systems do not include a strategy to deal with this kind of problem. Therefore, they are not robust in the presence of images generated by multiple sources.

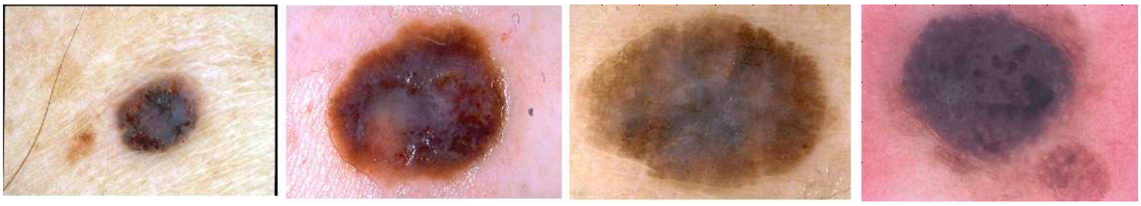


Figure 6.1: Multi-source examples. Images from [9].

Besides making it difficult to compare different methods, the lack of independence regarding the acquisition setup makes most of the CAD systems unsuitable to be used in clinical practice, since nowadays it is common to use teledermoscopy [143]. This is a technique where the dermoscopy images are acquired at local health facilities and sent to a central hospital, to be analyzed by a dermatologist. Each health unit is equipped with a different dermatoscope and acquire the images using specific illumination conditions. Since the CAD systems are not prepared to cope with this situation, their performance will be severely degraded. The same problem is observed when one uses the commercial EDRA database of dermoscopy images, which was acquired at three hospitals (University Federico II of Naples and University of Florence, both in Italy, and University of Graz in Austria) using different acquisition setups [9].

The goal of this chapter is to investigate strategies to deal with the problem of color-based CAD systems that are trained and tested using multi-source images. To deal with the multi-source problem it is assumed that color variations can be corrected using color constancy algorithms, which have been used in other pattern recognition and feature extraction problems [77].

6.2 Related Work

6.2.1 Color Normalization Applied to Dermoscopy Images

Different research groups have proposed color normalization strategies to deal with dermoscopy images. Most of the approaches are hardware-based [80, 83, 147, 195]. These approaches calibrate images by determining a set of internal camera parameters (e.g., camera offset, color gain and aperture) as well as a transformation matrix that is used to convert the images to a device independent color space.

Haeghen et al. [83] were among the first ones to propose a calibration model of this type. Their calibration procedure consists of converting the images from an unknown RGB color space, which depends on the acquisition system, to the device invariant sRGB space. Calibration is performed in a set of sequential steps. First they start by sequentially determining the specific parameters of the acquisition system, namely the camera offset, the frame grabber, the camera aperture, and its color gain. By knowing these four parameters it is possible to maximize the dynamic range and resolution of the system. Then, they use the 24 Gretag-Macbeth ColorChecker's (GMCC) patches to compute the parameters of the transformation matrix. This task is performed in two different stages. First, using a spectrophotometer, they acquire the CIE L*a*b* values of each patch after their conversion to sRGB. This allows them to determine the real values of the transformation from RGB to sRGB. Next, they compute the specific transformation matrix of the imaging system. This task is accomplished by acquiring the 24 GCMM patches using the imaging system. With these two sets of values, it is possible to obtain a set of linear equations that can be used to estimate the components of the transformation matrix.

Grana's et al. calibration model [80] starts with the correction of border and illumination defects. The former is applied to remove the black pixels associated with the frame of the image or with the black ring of the dermatoscope. The later consists of a filtering step, whose purpose is to correct the regions of the image where the illumination is not uniform. This filtering step is carried on separately for each color channel. The next step computes the gamma value of the camera and corrects it in all the images. With this correction, they obtain the RGB values that can be transformed into device-independent XYZ values. To determine the coefficients of the matrix that transforms RGB in XYZ they follow the same approach as Haeghen et al. [83], using the GCMM patches and XYZ instead of La*b*. Finally, they convert the images from XYZ to a new standard color space. Grana et al. state that the sRGB space used by Haeghen et al. [83] is not appropriate for dermoscopy images, since the color contrast is lower. Therefore, Grana et al. propose a new color space to describe the images. To determine the parameters of the conversion matrix from XYZ to the new space they have used a set of different colors extracted from dermoscopy images.

Wighton et al. [195] proposed a color calibration model for low-cost digital dermatoscopes. Their method not only corrects color and inconsistent illumination, but also deals with chromatic aberrations. First they start by performing color correction. This task is carried on as in the work of Grana et al. [80]. The following step is lighting calibration. Wighton et al. start by creating an illumination map for each channel of XYZ. This taks is performed using the white patch of the GCMM. After acquiring the patch, its XYZ values are compared with the ground truth values obtained with the spectrophotometer. The ratio between the ground truth and the acquired values lead to the correction maps. Finally, they correct the chromatic aberrations.

The main issue with hardware-based calibration methods is that they require the estimation of the device parameters as well as of the conversion matrix. In both cases, it is necessary to have access to the acquisition device in order to be able to work with the GMCC. It is difficult to obtain this kind of information when one is working with commercial databases such as EDRA [9] or when using teledermoscopy, since in this case the images are acquired at different clinical units. Furthermore, after a period of time the acquisition system requires re-calibration (*e.g.*, [83]), which might be time consuming and, consequently, overlooked.

To tackle the aforementioned issues, Iyatomi et al. [88] proposed a calibration system that is software-based. Their method performs a fully automated color normalization using image content in the HSV color space. Although Iyatomi's et al. method does not require knowledge about the acquisition setup, it has a training step. In this step, they start by extracting simple HSV color features from a dataset of dermoscopy images. Then, they use these features to build a set of independent normalization filters. In this stage, they include a selection process in which they reject the less relevant filters. The learning of the filters increases the conceptual implementation and complexities of the method. Furthermore, a filters' bank has to be learned whenever the training set changes.

This chapter explores a different approach to normalize dermoscopy images. The approach is based only on image information and does not required knowledge of the acquisition setup or a training step. Furthermore, it has been shown to significantly improve the performance of the CAD system when the images are acquired by multiple sources [13].

6.2.2 Other Color Normalization Algorithms

Color normalization has been studied by the computer vision and image processing communities. Among the different strategies proposed to normalize image colors there are approaches that try to account for the color of the light source, called color constancy algorithms [77, 101, 168]. A close examination of Figure 6.1 shows that part of the reason why the images look so different is the color of the light source. Particularly, the 4th image was clearly acquired using a light source that is saturated on the red channel. This evidence suggests that color constancy is suitable to deal with uncalibrated dermoscopy images.

One of the most popular color constancy methods is the gamut mapping described by Forsyth [71]. To be able to estimate the illuminant color, this method assumes that there is only a limited set of RGB values that can be observed under a given light source. This set is called *gamut* and the objective of the algorithm is to estimate the transformation that maps an observed gamut into the canonical one, where the canonical gamut is the set of RGB values observed under the canonical white light. Despite its potential high accuracy, gamut mapping shows some drawbacks: it requires training data acquired under a known light source and it is difficult to implement [77]. As a matter of fact, these two drawbacks are common to any of the other learning-based color constancy methods that can be found in literature, such as the color-by-correlation method proposed by Finlayson et al. [68] or the semantic-based method proposed by van de Weijer et al. [188] (see [77] for a description and comparison of different methods). Applying these methods to a dataset such as EDRA would be impractical since this would imply separating the images from different hospitals using their color content.

An alternative to the previous methods are the much simpler statistics based algorithms. These methods use low-level image features, like the mean value or the maximum response, to estimate the color of the light source. It has been demonstrated that, with appropriate parameter values, these methods achieve similar performance to that of more complex methods [68, 77, 188]. Furthermore, they are simple to implement, fast, and only require the tuning of a few parameters. This thesis considers four methods to compensate color distortions: gray world [30], max-RGB [105], shades of gray [69] and general gray world [188].

6.3 Color Constancy

6.3.1 Color Constancy Framework

The goal of color constancy is to transform the colors of a color image I, acquired using an unknown light source, so that they appear identical to colors under a canonical light source [77, 101, 168]. Usually, it is assumed that this canonical light source is the perfect white light. Color transformation is accomplished in two separate steps. First, the color of the light source is estimated in the RGB color space $[e_R\ e_G\ e_B]^T$. Then, the image is transformed using the estimated illuminant.

Different algorithms can be used to estimate the color of the illuminant. In this chapter four algorithms that use image statistics to estimate the color of the illuminant are investigated: gray world [30], max-RGB [105], shades of gray [69], and general gray world [188]. For a color image I, each component of the illuminant e_c, $c \in \{R, G, B\}$, is estimated based on the following expressions (see [77] for details)

- **Gray world**

$$\frac{\int I_c(\mathbf{x})\mathrm{d}\mathbf{x}}{\int \mathrm{d}\mathbf{x}} = ke_c \ \ , \tag{6.1}$$

- **max-RGB**

$$\max_{\mathbf{x}} I_c(\mathbf{x}) = ke_c \ \ , \tag{6.2}$$

- **Shades of gray**

$$\left(\frac{\int (I_c(\mathbf{x}))^p\mathrm{d}\mathbf{x}}{\int \mathrm{d}\mathbf{x}}\right)^{1/p} = ke_c \ \ , \tag{6.3}$$

- **General gray world**

$$\left(\frac{\int (I_c^\sigma(\mathbf{x}))^p\mathrm{d}\mathbf{x}}{\int \mathrm{d}\mathbf{x}}\right)^{1/p} = ke_c \ \ , \tag{6.4}$$

where I_c denotes the c-th component of image I, $\mathbf{x}= (x, y)$ denotes the pixel coordinates, and k is a normalization constant that ensures that $\mathbf{e} = [e_R\ e_G\ e_B]^T$ has unit length with respect to the Euclidean norm. Shades of gray and general gray world use the Minkowski norm to estimate the color of the illuminant, which depends on a parameter p, chosen by the user. Finally, $I_c^\sigma(\mathbf{x})$ is a smoothed image, obtained by filtering $I(\mathbf{x})$ with a Gaussian lowpass filter with standard deviation σ. Both σ and p can be tuned according to the dataset.

An interesting aspect of the studied color constancy algorithms is that they are related to each other. Gray world and max-RGB are special cases of shades of gray, for $p = 1$ and $p = \infty$, respectively. General gray world is an extension of shades of gray where image noise is removed by lowpass filtering.

After estimating **e**, it is possible to transform the image I. A simple way to model this transformation is by using the von Kries diagonal model [193]

$$\begin{pmatrix} I_R^t \\ I_G^t \\ I_B^t \end{pmatrix} = \begin{pmatrix} d_R & 0 & 0 \\ 0 & d_G & 0 \\ 0 & 0 & d_B \end{pmatrix} \begin{pmatrix} I_R \\ I_G \\ I_B \end{pmatrix} \quad , \tag{6.5}$$

where $[I_R, I_G, I_B]^T$ denotes the pixel value acquired under an unknown light source, and $[I_R^t, I_G^t, I_B^t]^T$ denotes the transformed pixel value, as it would appear under the canonical light source,. The canonical light source is assumed to be the perfect white light, *i.e.*, $\mathbf{e}^w = (1/\sqrt{3}, 1/\sqrt{3}, 1/\sqrt{3})^T$. The matrix coefficients $\{d_R, d_G, d_B\}$ are related to the estimated illuminant **e** as follows

$$d_c = \frac{1}{\sqrt{3}e_c}, \quad c \in \{R, G, B\} \quad . \tag{6.6}$$

6.3.2 Gamma Correction

Image acquisition systems, such as the ones used to acquire dermoscopy images, transform sRGB values using gamma (γ) correction curves, leading to what is called non-linear R'G'B'. In practice, this correction is applied for visualization purposes, since it reduces the dynamic range, *i.e.*, increases the low values and decreases the high values, as can be seen in (6.7), where $I_c(\mathbf{x}) \in [0, 1]$ and $c \in \{R, G, B\}$.

$$I'_c(\mathbf{x}) = \begin{cases} 12.92 I_c(\mathbf{x}), & \text{if } I_c(\mathbf{x}) \leq 0.0031308 \\ 1.055 I_c(\mathbf{x})^{1/\gamma} - 0.055, & \text{otherwise} \end{cases} \tag{6.7}$$

The main problem with γ correction is that digital systems store images after this correction is applied. However, the color constancy algorithms are derived for the sRGB values. Thus, before processing the dermoscopy images it is necessary to undo the γ correction. The transformation of R'G'B' to sRGB is simply performed by inverting (6.7) [145]

$$I_c(\mathbf{x}) = \begin{cases} \frac{I'_c(\mathbf{x})}{12.92}, & \text{if } I'_c(\mathbf{x}) \leq 0.03928 \\ \left(\frac{I'_c(\mathbf{x}) + 0.055}{1.055}\right)^{\gamma}, & \text{otherwise} \end{cases} \tag{6.8}$$

where γ is set to the standard value of 2.2 [145].

6.3.3 General Color Constancy Framework

The block diagram of the experiments carried on this chapter is displayed in Figure 6.2. First, gamma correction is performed using (6.8). Then the color of the lighting source is estimated using one of the methods described in Section 6.3.1and the image colors are normalized using (6.5). If one wants to use other color space besides RGB, the following step consists of transforming the color components of each pixel into the new color coordinates. Finally, the CAD system is applied to the image to diagnose it. Figure 6.3 exemplifies the color normalization process on an image from the EDRA dataset using the shades of gray method ($p = 6$).

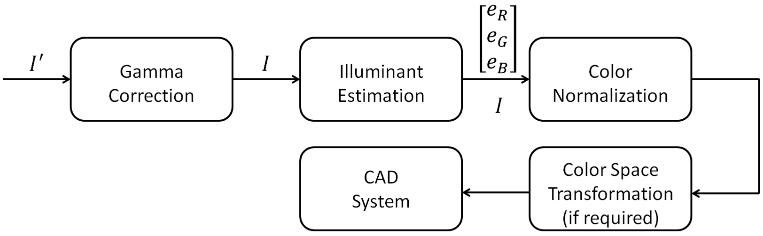

I'
Gamma Correction
I
Illuminant Estimation
$\begin{bmatrix} e_R \\ e_G \\ e_B \end{bmatrix}$
I
Color Normalization
CAD System
Color Space Transformation (if required)

Figure 6.2: Color normalization framework.

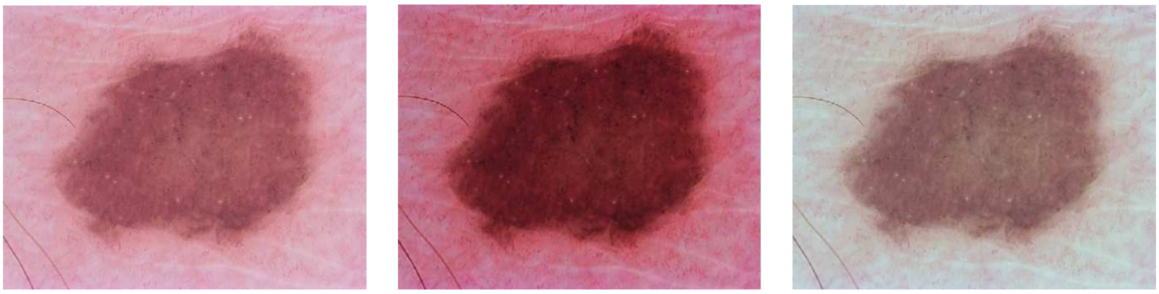


Figure 6.3: Example of color normalization: original image (left), gamma correction (middle), and corrected image using shades of gray ($p = 6$).

6.4 Experimental Framework

The BoF model is used to evaluate the performance of the color constancy methods. This algorithm was presented and evalutated in Chapter 5. The experimental pipeline is shown in Figure 6.4.

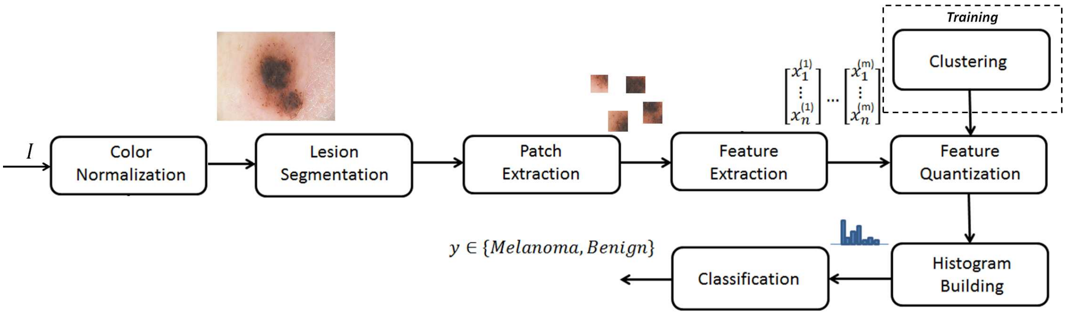

Training
Clustering
I
Color Normalization
Lesion Segmentation
Patch Extraction
Feature Extraction
$\begin{bmatrix} x_1^{(1)} \\ \vdots \\ x_n^{(1)} \end{bmatrix} \cdots \begin{bmatrix} x_1^{(m)} \\ \vdots \\ x_n^{(m)} \end{bmatrix}$
Feature Quantization
$y \in \{Melanoma, Benign\}$
Classification
Histogram Building

Figure 6.4: Block diagram of the experimental system. Image from [13].

The first step is to apply one of the color constancy algorithms described in the previous section. Then, the lesion is separated from the healthy skin. As before, the lesions were manually segmented by an expert. Image sampling is performed using the Harris-Laplace keypoint detector, which has been shown to perform well in dermoscopy image classification [18]. The support region of each keypoint is a square patch of size 40×40 centered on the keypoint. Patches that intersect the lesion in less than 50% of its area are excluded.

Image patches are characterized by 1-D color histograms. The selection of color features is directly related with the goal of this work, since the values of these features are the ones that are most influenced by changes in the acquisition setup. Two color spaces are used in the experiments. The first one is RGB, since this is the default color space of the dermoscopy images and it is highly dependent on the color of the light source. Experiments are also performed using the HSV space. Studies have shown that color constancy can also be applied before converting RGB images to other color spaces. Thus, this hypothesis is tested in this work using 1-D HSV histograms.

Clustering, feature quantization, and histogram building are performed as described in Chapter 5. Lesion classification is performed using a SVM classifier with the RBF kernel.

6.5 Experimental Results

6.5.1 Dataset and Evaluation Metrics

Previous chapters used the PH^2 database acquired under controlled conditions at a single hospital (Hospital Pedro Hispano, Matosinhos). In this chapter, all the experiments were carried on using the EDRA database [9] that contains images collected at three different hospitals: University Federico II of Naples (Italy), University of Graz (Austria) and University of Florence (Italy). Some examples of this database can be seen in Figure 6.1. A set of 482 images (50% melanomas) was selected as follows. First, most of the available melanomas were considered. Then, the same number of benign lesions was randomly selected from the following categories: blue nevi, clark nevi, spitz nevi,combined nevi, and dermal nevi.

Five BoF models were trained. The first was trained using non-normalized images and the remaining four were trained using one of the four previously described color constancy methods. A set of hyperparameters was tunned for each of the models using the grid search method. The hyperparamters optimized for all systems were the number of bins of the RGB/HSV histograms $\{4, 16, 24\}$, the number of centroids $\{25, 50, ..., 300\}$, the penalty C given to the soft margin in SVM $C \in \{2^{-5}, 2^{-4}, ..., 2^{5}\}$, and the width of the RBF kernel $\rho \in \{2^{-6}, 2^{-5}, ..., 2^{6}\}$. Specific parameters of the shades of gray (6.3) and general gray world (6.4) were also optimized: $p \in \{1, 2, .., 10\}$ and $\sigma \in \{1, 2, ..., 5\}$, as proposed in [77]. All the systems were evaluated using the SE and SP statistics. These metrics were computed using a 10-fold cross validation method. The final number of trained systems was 1081080.

6.5.2 Results

Figure 6.5 shows the values of the light source for each image of the dataset, estimated using the four previously described color constancy algorithms (see Section 6.3).

It is interesting to notice that the gray world (6.1) (1st row), shades of gray (6.3) (3rd row), and general gray world (6.4) (4th row) methods consider that most of the images need a significant cor-

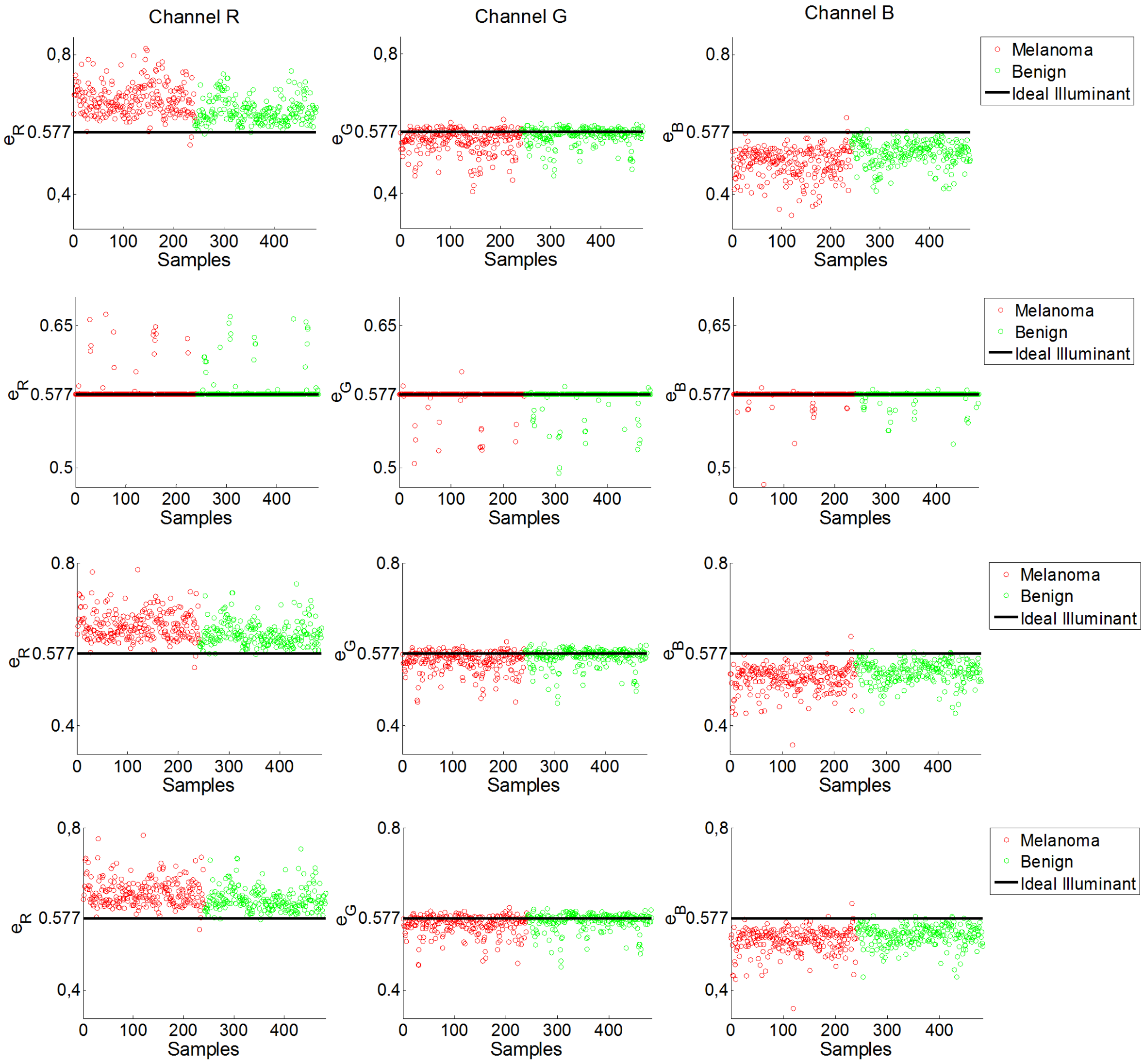


Figure 6.5: Estimation of the color of the light source for each of the images (samples) and each of the color channels using Gray World (1st row), max-RGB (2nd row), Shades of Gray $p = 6$ (3rd row) and General Gray World $p = 3$ and $\sigma = 2$ (4th row). The black line corresponds to the ideal color of the light source (white light source).

rection since there is a significant deviation between the estimated light source and the white one with $e_R = e_G = e_B = 1/\sqrt{3} \simeq 0.577$. On the other hand, according to the max-RGB method (6.2) (2nd row) there are less images that need to be normalized. This might be explained by that fact that while the three first methods use all the pixels in the image to compute the value of the illuminant in each color channel, max-RGB only uses the pixel with the highest value in each color channel.

Some examples of normalized images, as well as the estimated values of the corresponding light sources are shown in Figure 6.6. These results clearly show that the color constancy algorithms alter the appearance of the images, making them look more similar. This is more noticeable in images illuminated by a reddish light source (see Figure 6.6 1st and 4th columns), where the colors of the lesions and surrounding skin become much more distinguishable. Furthermore, color constancy also seems to enhance the contrast inside the lesion (see Figure 6.6 2nd column) and between the lesion and the surrounding skin, as can be seen in Figure 6.6 1st column. This last image is very interesting

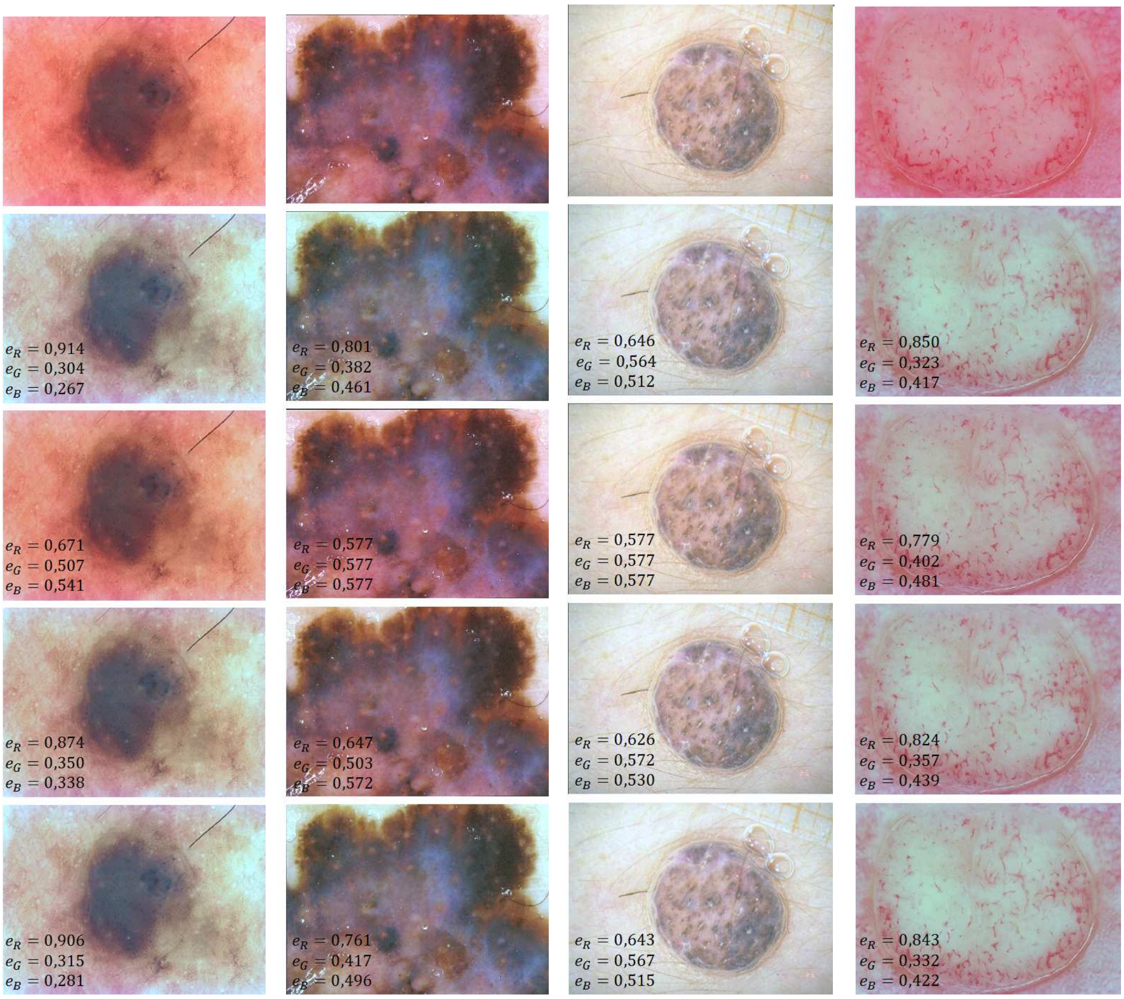


Figure 6.6: Color constancy examples. From top to bottom Original Image, Gray World, max-RGB, Shades of Gray ($p = 6$), and General Gray World ($p = 2,\ \sigma = 3$). Images from [13].

because without the color constancy it would be almost impossible to notice the light brown area that surrounds the darker center of the lesion, which might lead to an incorrect segmentation and consequent loss of color information. Separately analyzing each of the color constancy algorithms, it is possible to notice that gray world (2nd row) and general gray world (5th row, ρ = 2, σ = 3) are the ones that most alter the images, giving them a grayish color. Shades of gray (4th row, ρ = 6) mitigates this effect, giving the images a more normal coloration. Max-RGB seems to be the algorithm that least alters the aspect of the lesions. This was already observed in Figure 6.5. Recall that this algorithm is the limit of shades of gray for $\rho = \infty$, which explains its performance.

Table 6.1 shows the classification results obtained by the best configurations of the CAD system and using 1-D RGB histograms. These results show that all four classification algorithms significantly improve the performance of the classification system. Shades of gray seems to slightly outperform the other methods. This was also verified in the case of the HSV color space, where the use of shades of gray improved the results from $SE = 73.8\%$ and $SP = 76.8\%$ to $SE = 73.9\%$ and $SP = 80.1\%$. This means that color constancy can be used to improve the performance of color features based on the HSV space.

Table 6.1: Classification results with and without color constancy, using RGB histograms.

Algorithm	SE	SP
None	71.0%	55.2%
Gray world	78.8%	75.2%
max-RGB	79.2%	75.5%
Shades of gray	**79.7%**	**76.0%**
General gray world	79.6%	75.6%

It is important to assess if other features, such as texture, can be affected by the use of color constancy. To answer this question, color constancy was applied to systems trained using SIFT features. These features have already been used in the classification of dermoscopy images [18]. Table 6.2 shows the obtained results. As expected texture features are not significantly influenced by the use of color constancy, and there is even a marginal improvement in the performance. This improvement is mainly noticeable in the case of shades of gray.

Table 6.2: Classification results for SIFT features with and without color constancy.

Algorithm	SE	SP
None	78.3%	63.9%
Gray world	80.4%	62.7%
max-RGB	78.8%	64.8%
Shades of gray	**80.2%**	**65.6%**
General gray world	75.9%	67.3%

Color constancy was also applied to the PH^2 database. This dataset was acquired at a single hospital with the same experimental setup. It was concluded that no degradation of the system performance is observed [15].

6.6 Conclusions

This chapter investigates the application of color constancy to normalize the colors of dermoscopy images was investigated. The importance of color normalization and the performance of the four studied methods was assessed using a BoF model to classify dermoscopy images. The experiments were performed using images acquired at three different hospitals.

The results showed that the performance of the system significantly improves when color constancy is used. This was verified not only for color features of RGB and HSV color spaces, but also for texture features. The results also suggest that shades of gray is the most suitable method, among the four tested.

It has also been shown that color constancy operation does not degrade the performance of the system when the system is applied to images from a single source.

All of the investigated methods are very easy to apply and do not need any training, which means that it is not necessary to have any information about the acquisition setup used to acquired the images.

7

Feature Fusion

Contents

7.1 Motivation

The goal of the previous chapters is the comparison of different types of features, in order to assess their relevance and discriminative power. Three different problems were investigated: i) comparison between four types of global features (color, texture, shape, and symmetry), performed in Chapter 4; ii) comparison between global and local features (color and texture), performed in Chapter 5; and iii) extension to a more challenging multi-source problem, described in Chapter 6.

Each of the previous studies led to interesting conclusions, as well as to the formulation of new questions. An important question that remains to be addressed is: what can one gain from combining different types of features? Almost all of the CAD systems found in literature use more than one type of features to describe lesions, which suggests that combining the information will help improve the detection scores. However, combining different features (called feature fusion) is not a simple task. Dermoscopy works usually adopt the strategy of concatenating all of the features into a single feature vector (early fusion) [104]. However, there is no guarantee that this is a good approach. Works in other areas of application have shown that early fusion is not always a good choice to combine different types of features (*e.g.*, color and texture) [58, 177], and suggest the use of another approach called late fusion [104]. The idea of late fusion is to train a set of classifiers, each one depending on a different feature, and in a second step, the classifier outputs are combined.

The goals of this chapter are: i) to assess if it is possible to significantly improve the results by combining different features (global and local - color and texture); and ii) to provide a comparison between the two fusion strategies (early and late fusion). A detailed study of the early fusion of several types of features was recently discussed in [149]. However, late fusion has never been applied to dermoscopy image classification.

It is important to stress that the goal of this chapter is not to perform an extensive study on feature fusion, but only to study two simple strategies to combine different features and improve classification scores.

7.2 System Overview

Figure 7.1 shows the block diagram of a CAD system with early fusion of image features. Similarly to the previous work, each of the lesions was manually segmented, in order to separate them from the healthy skin.

As discussed in Chapters 4 and 5, different features are used to characterize the properties of skin lesions. These features can be separated into global features (representing the whole lesion by a single feature vector) and local features (obtained by dividing the lesion into smaller regions, each one characterized by a feature vector). Taking into consideration the results obtained in the previous chapters, both types of features will be used:

- **Global features:** two classes of global features are considered in this chapter, namely color histograms in three different color spaces: HSV, L*a*b*, and opponent, and texture (gradient's

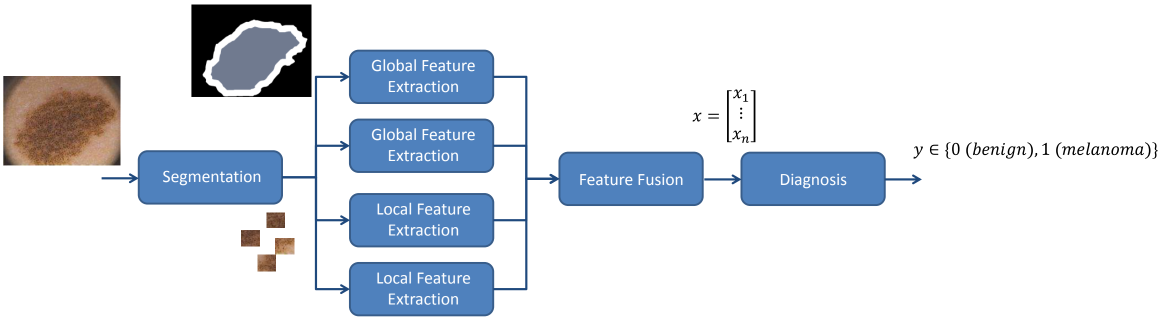

Segmentation
Global Feature Extraction
Global Feature Extraction
Local Feature Extraction
Local Feature Extraction
Feature Fusion
Diagnosis
$x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$
$y \in \{0\ (benign), 1\ (melanoma)\}$

Figure 7.1: CAD system using early fusion.

Table 7.1: Features and respective parameter values.

	Parameters	
Feature	Global	Local
Color Histograms - HSV (C1), L*a*b* (C2), and Opponent (C3) Spaces	32 bins per channel	16 bins per channel
Amplitude Histogram (T1)	16 bins	
Orientation Histogram (T2)	16 bins	
Gabor Filters (T3)	$N = 8$ orientations and 3 scales $\sigma_G \in \{2, 4, 8\}$	

orientation and amplitude histograms, and Gabor filters). All of these descriptors are computed as described in Chapter 4, *i.e.*, the lesion is divided into two regions (border and inner part), and features are separately extracted for each of these areas.

Shape and symmetry features are not used because it is not possible to computed these features when the lesion is not fully contained within the image borders. As discussed in Chapter 4, the application of these features leads to a reduction of the PH^2 dataset from 200 to 165 images, with only 12 melanomas out of the original 40. Therefore, in order to avoid excluding melanomas, shape and symmetry features were discarded.

- **Local features:** these features are computed using the BoF approach described in Chapter 5. The images are represented by a set of square patches of size 40×40 pixels, each of them characterized using color and texture descriptors. These descriptors are the same ones used as global features.

Table 7.1 shows a summary of the used color and texture descriptors, as well as the values of their hyperparameters.

The *feature fusion* block will be discussed in the next section. The *diagnosis* block classifies the lesion as melanoma or benign. This block is trained using a set of images previously diagnosed by a dermatologist. During the test phase, the learned classifier is applied to new images to predict their class. This will be done using the four classification algorithms considered in Chapters 4 and 5: AdaBoost [72], SVM [47] with a RBF kernel, kNN [60], and random forests [29].

7.3 Feature Fusion Strategies

This chapter aims to compare possible schemes for feature fusion. The approach used in most CAD systems is called early fusion, and consists of concatenating global features into a single feature vector. This is also a traditional approach in other research fields. However, several authors (*e.g.*, [58, 177]) question if early fusion is the most suitable approach when one is working with significantly different types features. Since one of the goals of this chapter is to incorporate the information provided by local features, early fusion may not be the best choice. Alternatively, several works use a methodology called late fusion, where the idea is to first train a set of classifiers, each one depending on a different type of feature, and then combine their outputs (*e.g.*, [135, 140]).

The investigation of feature fusion in the context of dermoscopy image analysis has never been performed, to the best of our knowledge. Thus, the study performed in this chapter may provide insightful information regarding this problem. The following sections describe the possible fusion schemes and point out their strong and weak points.

7.3.1 Early Fusion

Figure 7.1 shows the scheme of a CAD system that uses early fusion. In this approach, different feature vectors are extracted and combined into a single representation. The easiest strategy consists of simply concatenating different feature vectors into a single one [104], without any further processing besides normalizing all the features as in (4.20). Then, this new vector is fed to a classifier to either learn the decision rule (training phase) or to predict the diagnosis (test phase). Since this is the approach commonly used in dermoscopy image analysis, it will also be the one evaluated in this chapter.

An advantage of this strategy is that it performs the learning and classification phases only once. However, the feature vector that results from concatenating all of the features belongs to a high dimensional space. This might hamper the learning process (curse of dimensionality) and lead to the need of an additional feature selection step [104].

7.3.2 Late Fusion

Similarly to early fusion, this strategy also starts with the extraction of different features. Then, each type of feature is used to train a different classifier and the scores s of all of the classifiers (assumed to be in the interval $s \in [0, 1]$) are combined in order to yield a final diagnosis. Figure 7.2 shows the general scheme for late fusion. Notice that the *feature fusion* block in Figure 7.1 is replaced by a set of classifiers and a *scores combination* blocks.

Different strategies can be applied to combine the scores of the trained classifiers [104, 135, 140]. In this work two of them are compared:

- **Majority Voting:** This is a simple strategy where the scores of the different classifiers are set to be either 0 (benign) or 1 (melanoma). A final decision is performed by counting the number

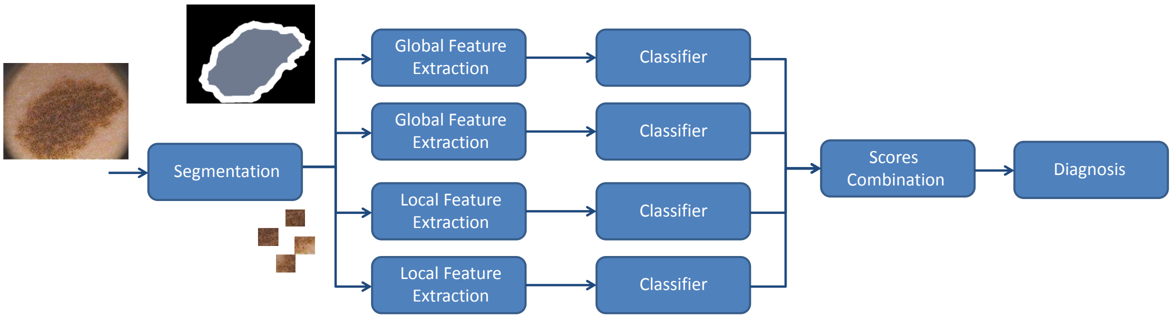


Figure 7.2: CAD system using early fusion.

of votes in each class, *i.e.*, the number of classifiers that gives 1 and those that give 0, and by selecting the class with the highest number of votes. This method requires an odd number of classifiers

- **Supervised Learning:** A more elaborate approach consists of combining the scores of different classifiers in a new feature vector that is fed to a final step of classification. This means that the training process comprises two stages of supervised learning: i) a set of classifiers is trained, each one using a different type of feature; ii) the scores of the classifiers are used to train a final classifier that predicts the diagnosis. During the test phase, the first set of classifiers produces the scores while the last classifier uses that info to obtain a final decision. In this chapter, the scores are used to estimate the parameters of a logistic regression.

The main strength of late fusion is the focus on the different performances of the features. However, this technique is expensive in terms of learning effort, since it is necessary to train several classifiers [104].

7.4 Experimental Results

7.4.1 Dataset and Evaluation Metrics

The CAD systems were implemented and tested using the PH^2 [126] and EDRA [9] databases. The first database is composed of 200 melanocytic lesions (40 melanomas), while the second contains more than 2000 images, including both melanocytic and non melanocytic lesions. A subset of images was selected from EDRA to carry on the experiments: 241 melanomas (most of the available ones) and 241 randomly selected benign melanocytic lesions among blue nevi, Clark nevi, Spitz nevi, combined nevi, and dermal nevi. All of the EDRA images were normalized using the methodology described in Chapter 6.

Different configurations were evaluated using both datasets. First, each type of features (global or local) as well as the different descriptors (the three color histograms and the texture features) were separately evaluated. Then, all of the possible feature fusion configurations were tested, ranging from combining only two features, to combining all of them. In all of the experiments, the trained systems were optimized by nested cross validation, in order to obtain the best possible generalization performance. Therefore, the parameters of the classifiers were searched in the range of values shown in Table 4.2, and the number of centroids K of the BoF algorithm was tuned according to $K \in \{100, 200, 300\}$.

The performance of each system is evaluated using SE and SP statistics, as well as the cost index (4.19). These metrics are computed using a nested 10-fold cross validation strategy, with hyperparameters selected by grid optimization.

7.4.2 Results

Four classification algorithms were tested in this chapter (Adaboost, kNN, SVM with a RBF kernel, and random forests). In order to keep the focus of the results on the comparison of fusion strategies, the results presented on this section are restricted to the best ones, obtained using random forests. Table 7.2 shows the best results for each type of feature and descriptor without fusion.

Table 7.2: Results for lesion diagnosis using single features and random forests. Best performance in bold.

		Global			**Local**		
Dataset	**Feature**	SE	SP	S	SE	SP	S
PH^2	C1	87%	86%	0.134	92%	79%	0.132
	C2	86%	86%	0.140	94%	77%	0.128
	C3	**89%**	**84%**	**0.130**	92%	78%	0.136
	T1	84%	80%	0.176	87%	85%	0.138
	T2	61%	63%	0.382	90%	79%	0.144
	T3	90%	63%	0.208	**88%**	**88%**	**0.120**
EDRA	C1	77%	69%	0.262	72%	65%	0.308
	C2	**79%**	**69%**	**0.250**	71%	66%	0.310
	C3	73%	72%	0.274	68%	69%	0.316
	T1	73%	56%	0.338	**82%**	**56%**	**0.284**
	T2	74%	54%	0.340	78%	56%	0.308
	T3	74%	62%	0.308	79%	55%	0.306

The top three best early fusion results, among the 4083 possible feature combinations, can be seen in Table 7.3. As expected, it is possible to improve the performance of the system by combining more than one type of feature. Interestingly, the best results for the EDRA dataset are achieved using only two types of features. It is also noteworthy that almost all of the setups where more than two features were combined led to worse performances. This can be a consequence of the high dimensional feature vectors that result from combining features using early fusion. Several works tackle this issue using a feature selection algorithm [103]. However, this strategy leads to the additional problem of defining which is the most suitable feature selection algorithm, which has not been extensively addressed in dermoscopy image analysis.

The top 3 best late fusion results out of 4083 possible feature combinations, can be seen in Table

Table 7.3: Lesion diagnosis results using early fusion and random forests. Top 3 best pair configuration/results - G stands for global feature and L stands for local feature. Best performance in bold.

		Results		
Dataset	Combination	SE	SP	S
PH^2	C2G + T1G	92%	89%	0.092
	C3G + T3L	94%	86%	0.092
	C2G + C3G + T2L +T3L	**98%**	**87%**	**0.068**
EDRA	C2G + T3L	77%	73%	0.243
	C3G + T3G	80%	70%	0.236
	C2G + T3G	**83%**	**68%**	**0.232**

7.4. These results were all obtained using supervised learning (logistic regression). Majority voting achieved worse scores, as can be seen in Table 7.5. The performance of late fusion is better than the one of early fusion. Moreover, it was possible to combine more than two types of features and significantly improve the classification scores (*e.g.*, see the EDRA results). Since late fusion consists of combining the outputs of different classifiers (recall Section 7.3.2), it allows the use of multiple descriptors without suffering from the curse of dimensionality that hampers early fusion. Another strength of late fusion (not investigated in this work) is the possibility of combining the outputs of different classifiers, *e.g.* combine the output of multiple SVM and random forests. This might improve even further the performance of a system, since one would be combining not only the strengths of different descriptors but also the power of several classifiers. The overall assessment of late fusion suggests that this method is the best strategy to be incorporated in a CAD system.

Table 7.4: Lesion diagnosis results using late fusion for supervised learning. Top 3 best pair configuration/results - G stands for global feature and L stands for local feature.

		Results		
Dataset	Combination	SE	SP	S
PH^2	C3G + T3G	93%	90%	0.082
	C1G + T1L + T3L	97%	88%	0.067
	C1G + C3G + T1L + T3L	**98%**	**90%**	**0.052**
EDRA	C1G + C2G + C3G + T3G +C2L + T2L + T3L	83%	72%	0.212
	C1G + C2G + C3G + T2L + T3L	81%	76%	0.208
	C1G + C2G + C3G + C2L + T2L + T3L	**83%**	**76%**	**0.198**

Examples of correctly classified lesions for both datasets can be seen in Figure 7.3. This images were correctly classified by the best late fusion systems highlighted in Table 7.4.

The main focus of this chapter was the improvement of the classification scores, and not to perform a comprehensive study on feature fusion. Therefore, the performed experiences were not exhaustive, in the sense that additional fusion strategies could have been considered. It is important to have in mind that feature fusion is a large field and that both early and late fusion can be performed in different

Table 7.5: Comparison of the best diagnosis results obtained using majority voting and supervised learning. Best pair configuration/results - G stands for global feature and L stands for local feature.

			Results		
Dataset	**Combination**	**Method**	SE	SP	S
PH^2	C1G + T1L + T3L	Majority voting	93%	95%	0.062
	C1G + C3G + T1L + T3L	Supervised learning	**98%**	**90%**	**0.052**
EDRA	C1G + C2G + C3G + T2L + T3L	Majority voting	83%	63%	0.250
	C1G + C2G + C3G + C2L + T2L + T3L	Supervised learning	**83%**	**76%**	**0.198**

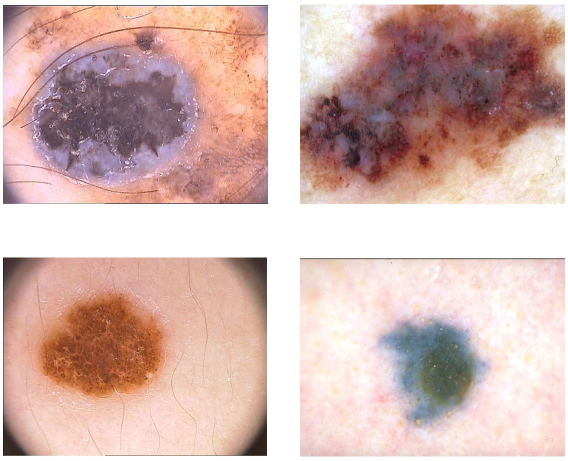


Figure 7.3: Correctly classified melanomas (top) and benign (bottom) lesions. Examples for the PH^2 (left) and EDRA (right) datasets.

ways. In the case of early fusion, the combination of several feature vectors into a single one does not have to be as simple as was performed in this chapter. An example is the early fusion strategy described in [135], where a kernel approach is used to combine the features. Alternatively, one can scale the different feature vectors, before concatenating them [58]. Similarly, different methodologies can be used to perform late fusion. Besides the methods described in this chapter, one could for example: i) compute the average, median, maximum or minimum value of the scores [100]; or ii) train a more sophisticated classifier to predict a diagnosis using the scores as features [104]. However, a thorough experimentation of all the possible methods was beyond the scope of this thesis.

7.5 Conclusions

This chapter addressed the problem of combining the different features (global and local) and descriptors (color and texture) that were independently studied in Chapters 4 and 5. Two feature fusion directions were explored: early and late fusion. The former is used in almost all dermoscopy works, while the latter had never been applied in the context of dermoscopy image analysis. Both strategies were studied in this chapter, in order to assess which of them was the most suitable.

The results showed that late fusion method is the best approach, with a $SE = 98\%$ and $SP = 90\%$ (PH^2), and $SE = 83\%$ and $SP = 76\%$ (EDRA), against $SE = 98\%$ and $SP = 87\%$ (PH^2) and

$SE = 83\%$ and $SP = 68\%$ (EDRA) obtained using early fusion. These results suggest that late fusion is the best approach to be incorporated into a CAD system. Nonetheless, additional options can be considered in both approaches (feature scaling, kernel methods, more complex classifiers).

8

Color Detection Using Gaussian Mixture Models

Contents

8.1 Motivation

Chapters 4, 5, and 7 address the detection of melanomas using pattern recognition based strategies. The first focused on the investigation of the role of different global features, namely, shape, symmetry, color, and texture, while the second tries to overcome the lack of a localized description, using a BoF model applied to local (color and texture) features. Finally, Chapter 7 discusses the fusion of multiple cues (color/texture, global/local). Both global and local feature extraction approaches are inspired by medical procedures. Global features are based on the ABCD rule of dermoscopy, while the local features are inspired by the localized analysis of patterns and colors performed by dermatologists. However, CAD systems that use these features are not easily accepted by dermatologists because they do not provide clinical information that justifies the diagnosis. This has been pointed out by different experts in [59]. Furthermore, the used features are general computer vision features that are often difficult to be understood by dermatologists and cannot be easily associated with medical cues.

An alternative strategy to overcome the previous issues is to develop a clinically oriented CAD system. The main characteristic of this type of system is the focus on the detection and characterization of dermoscopic criteria that are considered relevant by the medical experts (recall Section 2.3.2). An important characteristic assessed by dermatologists is the color of the lesion. This aspect is considered in different medical procedures, such as the ABCD rule [182] or the 7-point check list [7]. Moreover, dermatologists are aware that there are colors that are more common in malignant than in benign lesions [9]. This chapter describes a methodology for computing a statistical model to represent five clinically relevant colors (black, blue-gray, dark and light brown, and white).

8.2 Related Work

Detection of color related criteria has been previously investigated by different research groups [37,41,56,112,119,123,164]. The vast majority of the works focused on the detection of melanoma related dermoscopic structures, such as blue-whitish veil or regression areas [7]. One of the first studies in this field used decision trees to classify the pixels of an image as blue-whitish veil or not [37]. Then, the detected veil regions were used to classify the lesions as melanoma or benign. Madooei et al. [119] introduced alterations to the previous method, improving the computational cost. They also proposed a new approach in which the blue-whitish veil was detected by color matching. To achieve this goal, the authors created a color palette that contained samples of the colors that are usually associated with blue-whitish veil. This color palette was constructed using the Munsell space. Di Leo et a. [56] proposed a method to detect both blue-whitish veil and white regression areas. To perform this task they started by segmenting the lesion into different regions, using the two principal components obtained using PCA. Then, they classified the segmented regions using a regression tree. To describe each of the segmented regions they computed the mean and standard deviation values in different color spaces: RGB, HSI, and CIE L*u*v*.

Different shades of blue can be also associated with the diagnosis of a melanoma. Inspired by this medical characteristic, Lingala et al. [112] proposed an approach to identify multiple shades of blue in dermoscopy images. They started by segmenting different blue areas using fuzzy sets. Then, each of these areas was characterized using shape, color, and texture features. Finally, a SVM algorithm was used to classify the lesion as melanoma or benign.

Other research groups followed a different direction and focused on the quantization and/or identification of clinically relevant colors. Some groups performed color quantization using relative histogram approaches or clustering, and then used this information to classify the lesion as malignant or benign [41, 103]. It is undeniable that an approach that tries to account for the number of colors is medically inspired, since color quantization is also performed in the ABCD rule. However, this type of color quantization is performed without using medical information about the quantified colors, *i.e.*, a description of the clinically relevant colors is not provided to the system. This means that, for example, the clustering of the colors might lead to color representations that do not have a medical counterpart. In order to solve this problem some research groups performed color quantization using a restrict number of colors (usually those considered in the ABCD rule) [123, 164]. The clinically relevant colors are individually identified using a color matching approach, in which a color palette is used to describe the colors. The construction of the color palette is performed during a training step and is built based on color regions manually segmented by one or more experienced dermatologists.

The goal of this work is to identify the presence of clinically relevant colors [182] in dermoscopy images. For each color, it is desirable that the algorithm is capable of identifying whether the color is present or not, as well as to segment the regions in the lesion where that specific color can be found. This last requirement will allow the dermatologist to validate the output of the method.

The designed system is a clinically inspired one, thus medical knowledge was incorporated in the training process. Towards this aim, a dermatologist was required to identify and segment the colors in different dermoscopy images. Then, this information was used to estimate a Gaussian mixture model to represent each color. Gaussian mixtures have already been used with success in challenging skin related problems such skin detection in videos and pictures [99], and dermoscopy image analysis (*e.g.*, lesion segmentation [124, 186], and global pattern detection [160]). To the best of our knowledge, Gaussian mixtures models have not been applied to detect clinically relevant dermoscopy colors.

8.3 Proposed System

Figure 8.1 shows the block diagram of the proposed system, which can be divided into pre-processing, learning, and testing phases.

Each image is first pre-processed in order to remove misleading artifacts, namely skin hair and reflection pixels. This is performed using the pre-processing approach described in Chapter 3. Another important pre-processing step is color normalization. As shown in Chapter 6, image colors are strongly influenced by the acquisition light source. To tackle this issue, color normalization is per-

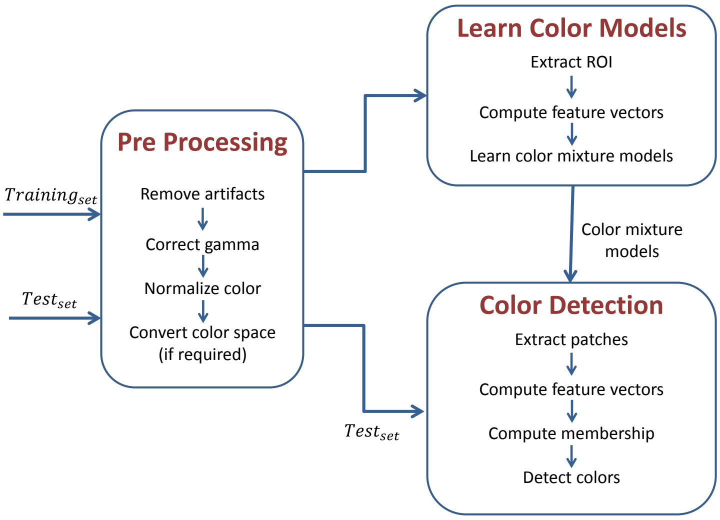


Figure 8.1: Block diagram of the color detection system.

formed using the Shades of Gray approach (6.3) with $p = 3$. Finally, the image can be converted to a different color space. Due to their properties, two color spaces are investigated in this chapter: RGB and HSV.

In this work it is assumed that each color can be modeled by a mixture of Gaussians. Each of the mixtures is learned using a set of patches, defined as region of interest (ROI) in Figure 8.1. These patches are extracted from color regions that were manually segmented by a dermatologist. A feature vector is used to characterize each patch and the parameters of the mixtures are learned using the algorithm proposed in [67].

The first step of the color detection block is to sample the lesion into 12×12 non-overlapping square patches and compute a feature vector to describe each of them. Then, the membership of each patch to the learned color mixtures is computed and a color label is assigned to it. Finally, the number of colors is estimated.

8.4 Learning Color Models

This section, describes the approach used to learn the Gaussian mixtures that represent the five colors.

8.4.1 ROI Extraction and Representation

Images from the PH^2 database are used to learn the color models. Among other characteristics, this database incorporates 29 images with medical segmentations of clinically relevant colors: dark brown, light brown, blue-gray, black, and white. Thus, these are the colors that are used in this work. For each of the 29 images, the different color regions were manually segmented and labeled by an expert dermatologist (see an example for each color in Figure 8.2 and consult [126] for further details). There are 17 examples of both dark brown and light brown, 6 examples of blue-gray, and 4 examples of both black and white. Only one example is available for the red color, thus this color is not considered.

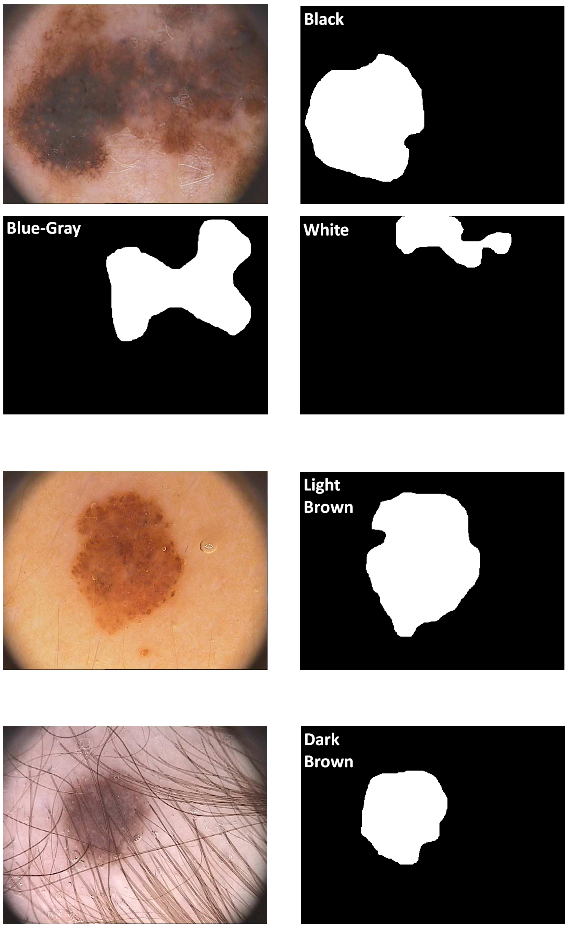


Figure 8.2: Examples of color regions medical segmentations.

A set of round training patches (ROIs), with a 5 pixels radius, was then randomly selected from each region. Since there are considerably more examples of dark and light brown than of the remaining colors, the number of patches extracted from each of the corresponding regions depends on the color. Therefore, 250 patches were selected from each light and dark brown regions, 350 patches from each blue-gray region, and 500 patches from each white and black regions. The final step computes a feature vector to characterize each patch, which is its mean color. Depending on the number of color spaces used, this feature vector can have a length of 3 or 6.

8.4.2 Learning Color Mixture Models

In this work, a statistical model is adopted to describe the colors, namely a Gaussian mixture model. Since the training set is composed of annotated data, it is possible to independently compute a mixture for each color. Thus, the final color palette comprises five different Gaussian mixtures, each with the following probability density function

$$p(\mathbf{y}|c, \theta^c) = \sum_{m=1}^{k_c} \alpha_m^c \; p(\mathbf{y}|c, \theta_m^c) \quad , \tag{8.1}$$

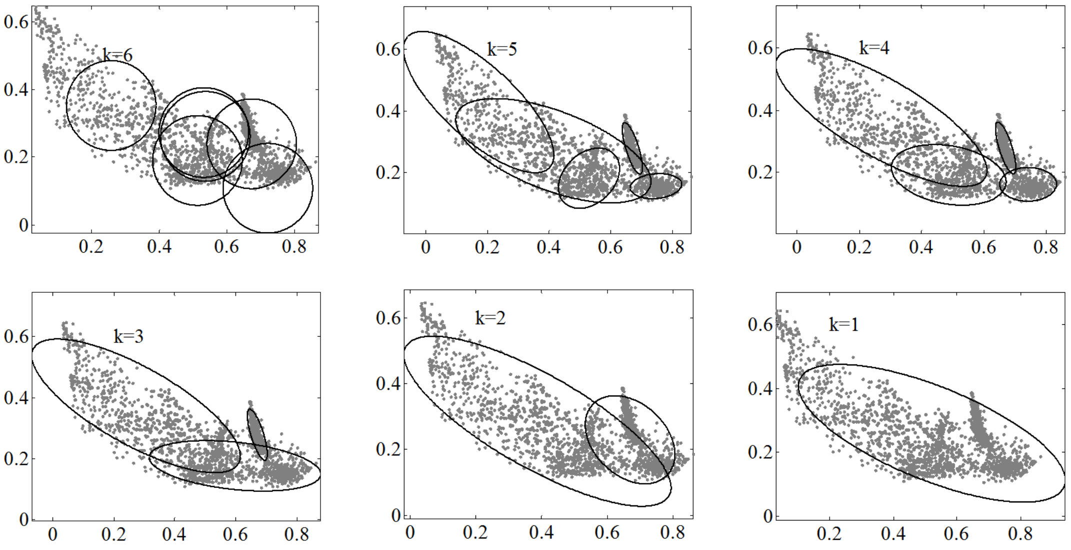


Figure 8.3: Fitting a Gaussian mixture to Black color examples. The gray dots are the projections of the RGB features on two of the dimensions and the black ellipses are level-curves of each component estimate. Initialization with $k_{max} = 6$. The algorithm selected $k_c = 4$.

where $c = 1, 2, ..., 5$ denotes one of the five colors, k_c is the number of components of the $c - th$ color mixture, $\alpha_1^c ... \alpha_{k_c}^c$ are the mixing probabilities ($\alpha_m^c \geq 0$ and $\sum_{m=1}^{k_c} \alpha_m^c = 1$), and θ_m^c is the set of parameters that defines the m-th component of the $c - th$ Gaussian mixture. In this work, **y** is a d-dimensional feature vector associated with each patch and

$$p(\mathbf{y}|c, \theta_m^c) = \frac{(2\pi)^{-\frac{d}{2}}}{\sqrt{|\Sigma_m^c|}} \exp\left\{-\frac{1}{2}(\mathbf{y} - \mu_m^c)^T (\Sigma_m^c)^{-1} (\mathbf{y} - \mu_m^c)\right\}, \tag{8.2}$$

where $\theta_m^c = (\mu_m^c, \Sigma_m^c, \alpha_m^c)$. Thus, the parameters to be estimated when learning a mixture are the mean and covariance matrix of each component $(\mu_m^c, |\Sigma_m^c)$, and the corresponding mixing probabilities α_m^c.

The most popular approach to estimate these parameters is the expectation-maximization (EM) algorithm. However, it is not easy to select the best number of components of the mixtures using this algorithm. To deal with this issue, Figueiredo and Jain [67] proposed an approach to learn mixture models, using a variant of the EM algorithm that implements the minimum message length (MML) criterion as the cost function, in order to automatically find the best number of components of a mixture. The procedure tests all the component numbers in $\{k_{min}, k_{min} + 1, ..., k_{max} - 1, k_{max}\}$ using the component-wise EM (CEM) for mixtures [42]. After achieving convergence with a certain k, they set the component with smaller $\widehat{\alpha}_m$ to zero and rerun CEM until convergence. This task is performed while $k \geq k_{min}$ and, in the end, the estimated parameters as well as the number of components are those which minimize the MML criterion. Furthermore, this algorithm is also able to overcome two other drawbacks of the EM algorithm: its dependency on initialization and the risk of convergence to the boundary of the parameter space, which may lead to meaningless results. For a detailed description, see [67].

An example of the results of the algorithm can be seen in Figure 8.3, where a Gaussian mixture is estimated for the training examples of the black color. In this example the algorithm was initialized with $k_{max} = 6$ and selected $k_c = 4$ as the best number of components. It is import to stress that the optimal number of components is not the same for all the colors. The training vectors **y** and the learned mixture parameters are d-dimensional (only the first two coordinates are shown in the figure).

8.5 Color Identification

A hierarchical decision scheme is adopted to identify the colors: patch labeling and lesion labeling. First, the lesions are separated from the surrounding skin, using manual segmentations performed by an expert. Then, each lesion is sampled into small patches of size 12×12 using a regular grid. This size was selected based on the average resolution of the dermoscopy images (570 × 760). Furthermore, these dimensions make it possible to identify small color regions in the lesions, without significantly increasing the computational running times. Finally, a feature vector is computed to characterize each patch, using the mean color as before. To assign a color label to each of the patches, the posterior probabilities of each color c are computed as follows

$$p(c|\mathbf{y}) = \frac{p(\mathbf{y}|c, \widehat{\theta}^c)p(c)}{p(\mathbf{y}|\widehat{\theta})} \quad , \tag{8.3}$$

where $\widehat{\theta}^c = (\widehat{\mu}^c, \widehat{R}^c, \widehat{\alpha}^c)$, $\widehat{\theta} = (\widehat{\theta}^1, ..., \widehat{\theta}^5)$, $p(c) = 1/5$ is set to be equal for all colors, and

$$p(\mathbf{y}|\widehat{\theta}) = \sum_{c=1}^{5} p(\mathbf{y}|c, \widehat{\theta}^c)p(c) \quad . \tag{8.4}$$

Then, the degrees of membership are sorted and the colors with the highest and second highest values are denoted as $c_1$and c_2, respectively. This information is used to either label the patch or reject it, as follows

1. If $p(c_1|\mathbf{y}) \geq \delta$ and $p(c_2|\mathbf{y}) < \epsilon p(c_1|\mathbf{y})$, where $\delta = 0.35$ and $\epsilon = 0.8$ are thresholds that have been experimentally determined, the patch is labeled according to color c_1.

2. If $p(c_1|\mathbf{y}) \geq \delta$ and $p(c_2|\mathbf{y}) \geq \epsilon p(c_1|\mathbf{y})$, the patch receives a label which expresses doubt between c_1 and c_2.

3. If $p(c_1|\mathbf{y}) < \delta$, the patch is rejected.

The final step consists of deciding whether a color is present or absent in a lesion. This task is performed using the patches previously labeled with one of the five colors. Patches that have doubt labels are not considered in this process. For each color, the area of the patches with the corresponding label (*i.e.*, the number of pixels) is computed and compared with an empirically determined area ratio threshold. Each color is validated only if its area ratio is above a specific threshold, where the area ratio (λ_c) for color c is defined as

$$\lambda_c = \frac{A_{patches^c}}{A_{lesion}} \quad . \tag{8.5}$$

$A_{patches^c}$ is the total area of the patches labeled with color c and A_{lesion} is the area of the lesion, computed using the lesion segmentation mask provided by an expert. A different area ratio threshold was experimentally determined for each color.

8.6 Experimental Results

8.6.1 Dataset and Evaluation Metrics

As described in Section 8.4, 29 images from PH^2 are used to learn the color models. These are the only images for which color region segmentations performed by an expert are currently available. A second set of 123 images is then used to validate the models. These images are different from the ones used for training and color region segmentations are not available. Nonetheless, each of the 123 images has a label stating whether each color is present or absent.

A robustness experiment is performed using 340 images from the EDRA dataset. Although color region segmentations are not available, each of the images is labeled according to the presence or absence of the colors. EDRA is a more challenging dataset, not only because it has a higher number of images, with greater color variability, but also because it is a multi-source dataset. Thus, the goal of this experiment is to assess the robustness of the previously learned color models and determine if they can be applied to other datasets acquired under different conditions.

To evaluate the performance of the algorithm, its outputs are compared with the labels provided by the dermatologists. This means that the lesion labels provided by the algorithm using (8.5) are compared with those of the expert. Then, three different statistics are computed:

- **Sensitivity (SE)** - Percentage of images for which each color was correctly identified;
- **Specificity (SP)** - Percentage of images for which each color was correctly non detected;
- **Accuracy (ACC)** - Percentage of correct decisions.

These statistics are computed for each of the colors, using a formulation of one-vs-all. Table 8.1 summarizes the sizes of the training and test sets, as well as the number of images for each color.

Table 8.1: Number of color labels per class of lesion.

		Color				
Type of set	**Database/#Images**	Dark Brown	Light Brown	Blue-Gray	Black	White
Training set	PH^2(#29)	17	17	6	4	4
Test set	PH^2(#123)	78	78	26	28	7
	EDRA(#344)	303	247	226	179	15

8.6.2 Results

Four different questions are addressed in this section:

i) Compare the performance of single Gaussian models against Gaussian mixture models;

ii) Compare the performance of the classical expectation maximization (EM) algorithm against the algorithm used in this work [67];

iii) Compare the performances of RGB only, HSV only, and RGB+HSV;

iv) Determine if the estimated models can be applied to other datasets acquired under different conditions.

The results for the first three questions were obtained using the PH^2 test set, while the results for the last question were obtained using the EDRA test set (see Table 8.1). All of the models were trained using the training PH^2 training set of 29 images, described in Table 8.1.

The experimental procedures used to solve the three first questions were the following. First, the five color models were learned using different configurations of feature vectors: RGB only, HSV only, and the combination of the two. In the case of RGB four different systems were computed for each color: single Gaussian, Gaussian mixtures of size 3/6/9 using EM, and Gaussian mixtures computed using [67]. For the later, the mixture learning algorithm was initialized with $k_{max} = 9$ and $k_{min} = 1$. The initial values of the mixture components were computed as described in [67].

Table 8.2 shows the average color detection scores using the four RGB systems. Gaussian mixtures outperform a single Gaussian, as expected. The procedure used in this work to estimate the mixture parameters as well as the number of Gaussians per color [67] shows a better average performance than classical EM. Although the system trained with EM and 6 Gaussians per color achieved a similar SE and a lower standard deviation, the average SP and ACC show significantly higher standard deviations. Method [67] achieves low standard deviations for the three statistics, meaning that it is more robust. The individual results for each color can be seen in Table8.3.

Table 8.2: Average color detection results on the PH^2 test set using RGB features.

Estimation Method	Average SE	Average SP	Average ACC
Single Gaussian	64.3%±21.7%	78.3%±8.9%	71.3%±10.3%
EM-3 Gaussians	67.4%±24.3%	76.4%±14.8%	71.9%±11.1%
EM-6 Gaussians	86.3%±3.5%	74.0%±16.8%	80.2%±8.5%
EM-9 Gaussians	82.4%±14.5%	70.3%±19.5%	76.4%±10.1%
Figueiredo and Jain [67]	**86.6%±6.4%**	**75.1%±3.8%**	**80.8%±3.7%**

Table 8.3 show the color detection results for the HSV space. When compared with the statistics of the RGB space, it seems like HSV leads to a marginal improvement of the average performance. HSV also seems more appropriate to describe the blue gray color. The best detection scores are achieved with the combination of the RGB and HSV feature vectors. Table 8.3 also shows the statistics obtained with this system. By combining the two color spaces it is possible to improve the overall performance of the system, achieving a better balance between SE and SP, and ACCs above 80% for all the color models.

Figure 8.4 shows some examples of the output of the algorithm for different lesions of the PH^2 test set, as well as their ground truth labels. The color normalized images were also included in order to exemplify how this process increases the color similarities among different images. It is interesting to notice that even in the bottom case, where the image has an overall reddish hue, the proposed

Table 8.3: Color detection results using RGB, HSV, and RGB+HSV.

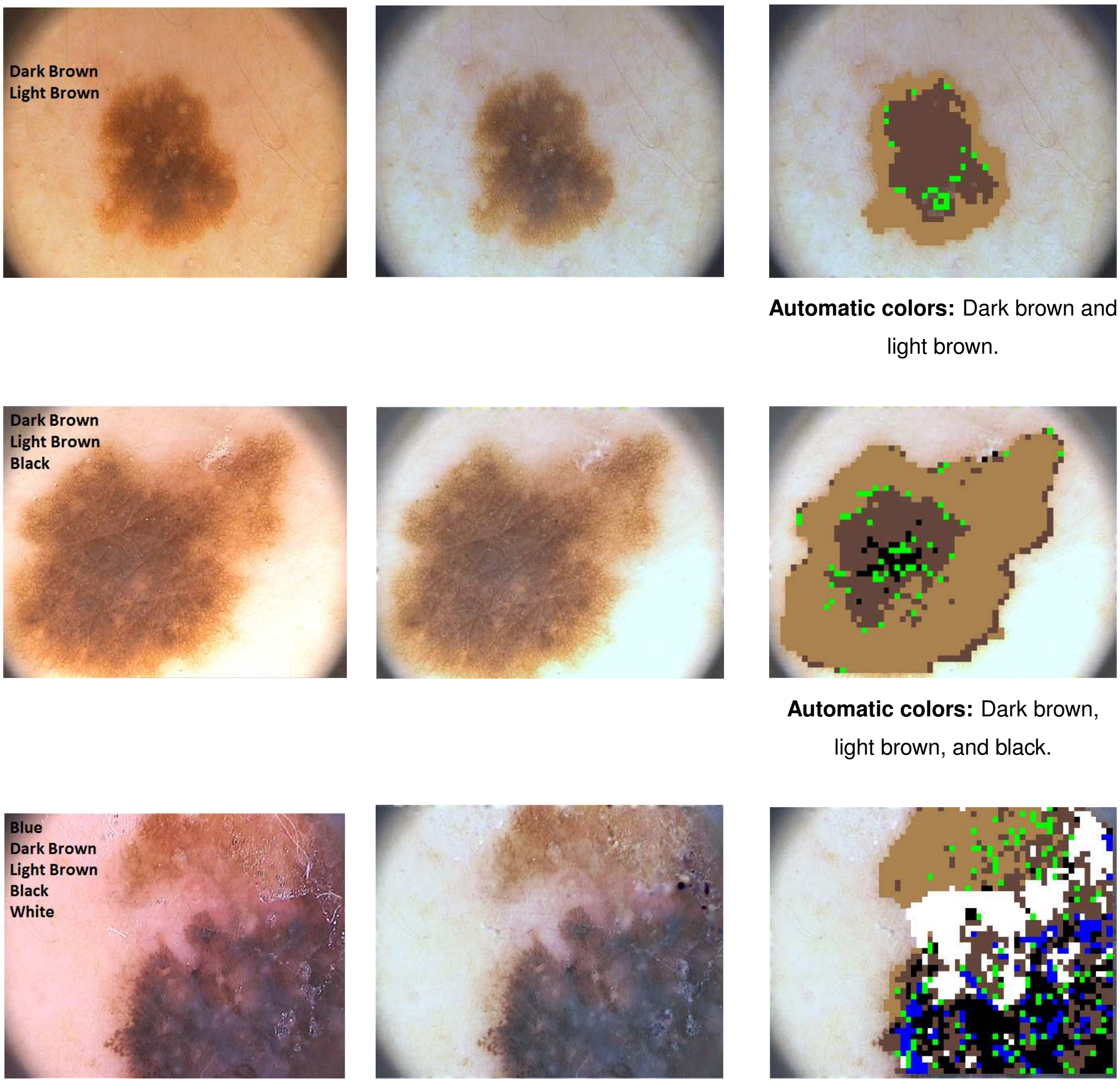
Color	**Color Space**	SE	SP	ACC
Blue-Gray	RGB	76.9%	74.0%	75.5%
	HSV	92.3%	80.5%	86.4%
	RGB+HSV	88.5%	83.1%	85.8%
Dark-Brown	RGB	93.6%	76.0%	84.8%
	HSV	94.9%	64.0%	79.4%
	RGB+HSV	92.3%	72.0%	82.2%
Light-Brown	RGB	91.0%	72.0%	81.8%
	HSV	91.0%	76.0%	83.5%
	RGB+HSV	92.3%	76.0%	84.2%
Black	RGB	85.7%	72.0%	78.9%
	HSV	85.7%	72.0%	78.9%
	RGB+HSV	92.9%	78.7%	80.4%
White	RGB	85.7%	81.3%	83.5%
	HSV	85.7%	74.0%	79.8%
	RGB+HSV	85.7%	78.1%	81.9%
Average	RGB	86.6%±6.4%	75.1%±3.8%	80.8%±3.7%
	HSV	89.9%±4.1%	73.3%±6.1%	81.6%±3.2%
	RGB+HSV	**90.3%±3.1%**	**77.6%±4.1%**	**84.0%±1.9%**

system is able to identify the colors in the lesion. The patches that were labeled as ”doubts” are also shown in Figure 8.4. A search for the most frequent doubt labels revealed that there are four that are more common than the others: the blue gray-black, the blue gray-dark brown, the dark brown-black, and the dark brown-light brown.

The best color models learned using the 29 images of PH^2 (RGB+HSV) were applied to the 340 images of the EDRA database. The color identification results can be seen in Table 8.4. As in the case of PH^2, the system achieved an overall good performance with an average ACC of 76.5%. The system performed well for most of the colors, especially if one considers that the number of test examples for each color has been significantly increased. Increasing the number of examples leads to higher color variability, which may justify the reduced performance of the system for the black color.

Table 8.4: Color detection results for the EDRA dataset using RGB+HSV.

	SE	SP	ACC
Blue-Gray	74.9%	81.0%	77.9%
Dark-Brown	81.7%	71.8%	76.8%
Light-Brown	83.8%	77.0%	80.4%
Black	70.3%	63.9%	67.1%
White	85.7%	76.5%	81.1%
Average	78.6%±9.2%	74.3%±6.5%	76.5%±5.6



Automatic colors: Dark brown and light brown.

Automatic colors: Dark brown, light brown, and black.

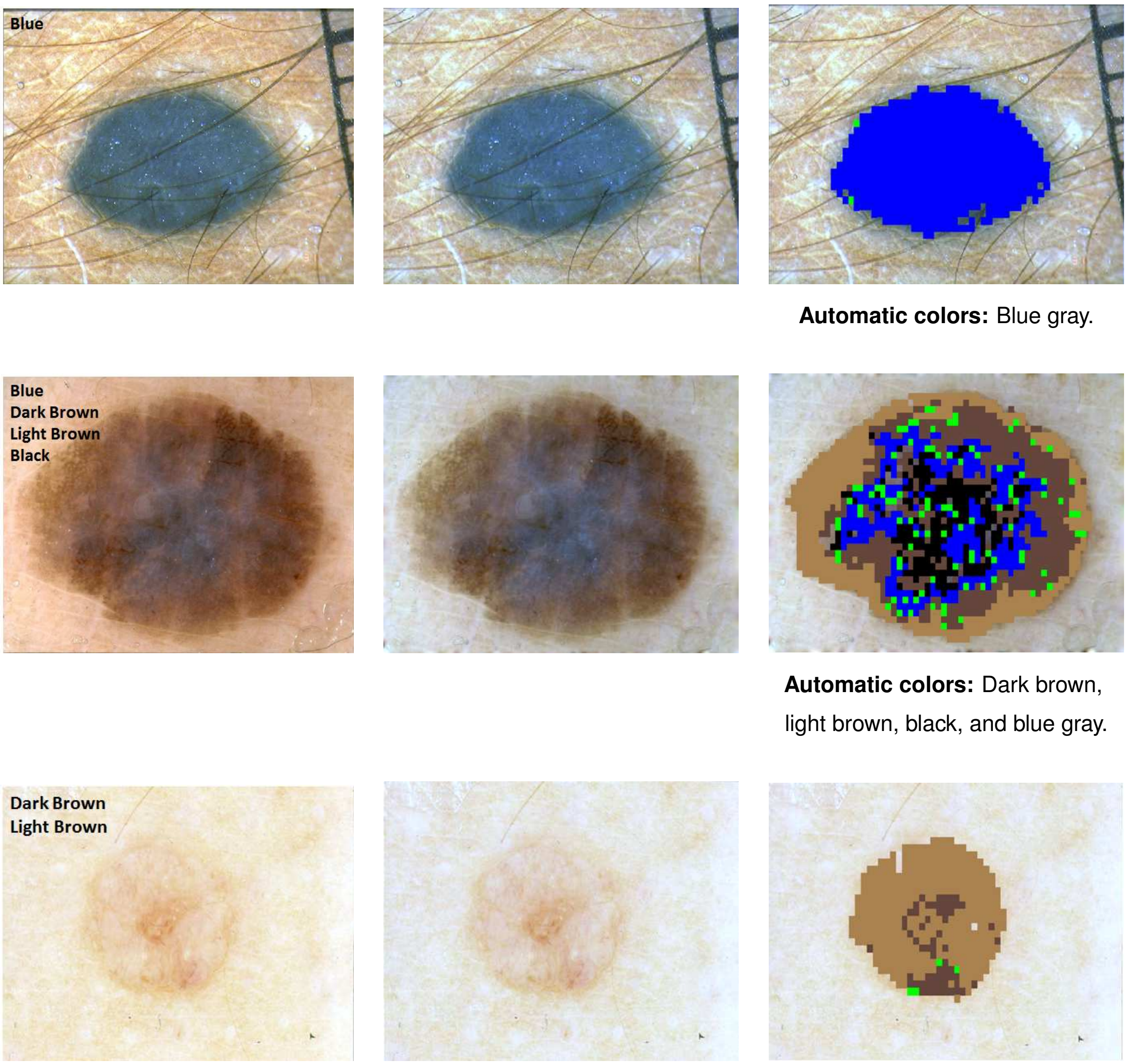
Automatic colors: Dark brown, light brown, black, blue gray, and white.

Figure 8.4: PH ^{2}examples : Original image and ground truth labels (left); Image after color normalization (middle), and Color Labels (right). The colored regions represent the identified colors labels, while the green pixels correspond to patches that have received one of the doubt labels.

Figure 8.5 shows some examples of the output of the system for the EDRA dataset. An interesting aspect about the example in the first row is that there is no blue-gray training region with a color similar to that of the lesion. Nonetheless, the system is still capable of correctly identifying that color as blue-gray.

8.7 Conclusions

This chapter described an approach for the identification of clinically relevant colors in dermoscopy images. Each of the clinical colors (black, blue-gray, light and dark browns, and white) were modeled using Gaussian mixtures. The Gaussian models were trained using 29 medically segmented images



Automatic colors: Blue gray.

Automatic colors: Dark brown, light brown, black, and blue gray.

Automatic colors: Dark brown and light brown.

Figure 8.5: EDRA examples : Original image and ground truth labels (left); Pre-processed image, and Color Labels (right). The colored regions represent the identified colors labels, while the green pixels correspond to patches that have received one of the doubt labels.

from the PH^2 database and were validated using two different sets: 123 images from PH^2, different from the ones used for training, and 340 images from the EDRA dataset. Two color spaces were investigated (RGB and HSV) and the experimental results showed that the best detection results are achieved by combining their information. The results were promising with and average accuracy of 84.2% on the 123 PH^2 images, and 76.5% on the EDRA dataset.

Despite the promising results, there is a significant difference in the performance of some of the color models in the two datasets. This might be related with differences color perception between the dermatologist who annotated the training set and those that annotated the EDRA dataset. Another explanation is the lack of similar examples on the training set. Both issues should be addressed in future work, but this is a difficult direction. On one hand, the number of examples on the training set should be increased but, as stated before, medical segmentations are difficult to obtain. This is a slow process, and expert dermatologists are not willing to segment clinical criteria because it is a subjective task. On the other hand, the color region segmentations should be performed by more than one expert. It would also be interesting to include training images acquired using different setups on the training set, since it may increase the robustness of the models.

The colors addressed in this chapter are a subset of the ones considered in the ABCD rule (see Table 2.2). This method assesses the lesion for the presence of six color, but unfortunately it was not possible to model the missing color (red/pink) due to lack of training examples for this color. The need of region segmentations to learn the color models can be seen as the downside of the proposed method. Nonetheless, it is important to stress that whenever enough training examples are provided, the algorithm is capable of learning color models and performs well, as shown by the results.

9

Towards the Development of a Clinically Inspired System

Contents

9.1 Motivation

The previous chapter described a methodology to detect the ABCD rule's colors. This method could be combined with the one described in Chapter 3, in order to develop a clinically inspired system that would perform the automatic diagnosis based on two medical criteria: colors and pigment network. However, dermatologists take into account more than two dermoscopic criteria when performing a diagnosis [7, 182], which means that it is necessary to increase the number of criteria detected by the system. As one might imagine, the development of several detection strategies, each one specialized for one of the dermoscopic criteria, is a cumbersome task. On one hand, some of the criteria are subtle structures that are not easy to identify and detect. On the other hand, the development of detection systems requires large datasets of images with detailed information: text annotations stating which are the dermoscopic criteria that can be found in the lesions, and their corresponding segmentations. It is very hard to find datasets that contain all of this information. Usually, the available ones lack the segmentations of the criteria, since this is seen as a subjective and time consuming task that dermatologists are not willing to do. The absence of segmentations for clinical criteria, hampers the development of CAD systems based on the independent models for each clinical criteria. An example is the color detection method proposed in Chapter 8, where it was not possible to learn a mixture model for the red/pink color due to lack of training examples.

This chapter proposes a strategy to detect dermoscopic criteria, which is able to deal with the aforementioned problems. Moreover, the detected structures are used to develop a diagnosis system that is able to distinguish between malignant and benign lesions. The proposed system is one of the first of its kind and, to the best of our knowledge, the framework used to detect the dermoscopic structures has never been applied to dermoscopy.

9.2 Related Work

9.2.1 Clinically Inspired CAD Systems

Clinically inspired CAD systems can be divided into two different categories: systems that are based on the pattern analysis method proposed by Pehamberger et al. [141] and systems that are based on the ABCD [182] or 7-point checklist [7] methods. An assessment of several works that fall in these two categories can be found in Section 2.3.2.

A common trait to several of the clinically inspired systems found in literature is the use of segmented dermoscopic criteria (*e.g.*, colors) to train a detection algorithm. Several pattern analysis methods use small portions of the lesions associated with the different patterns considered in [141]. Each of these regions is characterized using texture descriptors, *e.g.*, filters [156] or a Markov random field [167]. Finally, this information can be used to learn a set of models, such as Gaussian mixtures [160] or a dictionary of textons [156], to represent each of the patterns. The analysis of a new image consists of finding the pattern template that better represents the features, and label the lesion accordingly.

Other kind of systems that require segmentations are the ones that try to automatically detect one or more of the six ABCD rule's colors [112, 123, 142, 161, 164]. The main step of these methods is the estimation of a color pallet to represent the spectrum of admissible colors in a lesion. This requires a representative training set, which is usually obtained by asking experienced dermatologists to segment those colors in a (small) set of dermoscopy images. As before, the new images are analyzed by matching their pixels with the templates. Works that focus on the detection of color structures, such as blue-whitish veil [37, 56, 57, 119] and regression areas [49, 54, 56, 57, 181], also require segmentations to train their models. Some of these methods extract features from positive and negative regions and use them to train a classifier, namely decision trees [37, 54, 56, 57] and neural networks [49, 181]. An alternative strategy consists of learning a color palette using the region examples and then use a nearest neighbor approach to label new regions according to the estimated palette [119].

The aforementioned works heavily rely on accurate segmentations of the different criteria performed by experts. Unfortunately, this kind of information is difficult to obtain, as dermatologists often provide text labels stating which are the dermoscopic criteria that they can observe in the lesions, but avoid segmenting them. This makes it impossible to reproduce a significant number of clinically inspired systems. Moreover, the lack of segmentations hampers the development of full systems, as happened in Chapter 8, where it was not possible to estimate a model for the red/pink color. Recently, Madooei et al. [120] attempted to deal with the lack of segmentations, using a multiple instance learning (MIL) framework to detect blue-whitish veil. The MIL algorithm belongs to the category of methods that are trained using weakly annotated data, *i.e.*, the ground truth is formed by text labels only. Thus, this algorithm is appropriate to cope with the lack of segmentations in dermoscopy image analysis. Despite the promising results, this method has not been extended to other dermoscopic criteria, nor has it been used to perform lesion diagnosis. This work was published at the same time as the first results of this chapter [12].

Although the need for accurate segmentations is clearly the main problem of the clinically inspired systems, other issues can be identified. Few works try to diagnose the lesions using the detected dermoscopic criteria (some exceptions are [5, 23, 37, 49, 57, 91, 112, 160, 164]), and even fewer attempt the detection of more than one criterion [49, 54, 56, 57, 142]. Lesion classification should be a critical test, since it is the best strategy to determine the discriminative power of the detected criteria. Nonetheless, performing the diagnosis using only one criterion is insufficient, since expert dermatologists base their decision on rules that consider a set of criteria.

This chapter describes a framework to deal with the aforementioned problems. The proposed CAD system is able to identify multiple dermoscopic criteria, using an image annotation algorithm that is able to learn from weakly annotated data (only text labels). The information provided by the criteria is then used to diagnose the lesions as malignant or benign.

9.2.2 Image Annotation

The automatic reproduction of human text labels is the goal of image annotation algorithms. This task can be seen as providing a full caption for an image or video, stating which are the most relevant concepts, and on some occasions complementing those concepts with specific regions of the image (semantic segmentation). One of the major challenges of image annotation algorithms is that they have to be trained using weakly labeled data, *i.e.*, they have image labels but no indication of the image regions that are connected to each of the labels [205, 209]. Recalling the previous section, it is possible to notice that this chapter faces a similar issue: there is a set of medical text labels to describe the lesions but the corresponding segmentations are missing. Thus, it makes sense to see the detection of dermoscopic criteria as an image annotation problem.

Different types of annotation algorithms have been proposed [86]. A popular strategy is to treat the annotation problem as a multi-class classification problem [209], which can lead to different formulations. The simplest one consists of decomposing the multi-label learning problem into several binary classification problems (one for each label) [27, 44, 208]. This approach has been shown to obtain interesting results. However, learning a separate classifier for each of the labels might not be practical if there is significant number of possible labels and training images. Furthermore, these methods ignore the existence of label correlations, which can lead to suboptimal results. More sophisticated methods address these issues either taking into account the correlation between any pair or group of labels [43, 76, 78, 94, 113, 146, 150, 198, 206, 211] or by converting the problem into one of label ranking [62, 74, 162, 207]. The main downside of these methods is their high computational complexity.

The classification-based strategies found in literature can also differ in the way they characterize the images. It is possible to either describe the whole image using a single feature vector, or to divide it in different regions and separately characterize each of them. In the later, an image receives a certain label if at least one of the regions is associated with it. These kind of methods are called multi-instance multi-label (MIML) learning methods [34, 63, 86, 110, 136, 191, 199, 203, 210]. One of the relevant aspects of MIML methods is that they are able to model the relationship between image labels and regions. However, the MIML framework has a set of drawbacks. It lacks robustness in the presence of outliers (a single outlier can bias the solution) and it is highly sensitive to initialization. Moreover, these methods usually require the definition of the number of instances, which can be accomplished by either breaking the images into a fixed number of regions or by applying the BoF framework before the learning phase. If the later is performed, the clustering step might introduce errors and create misleading prototypes. Finally, they rely on decomposing the learning task into a series of single class multi-instance learning procedures, *i.e.*, into multiple independent binary classification problems.

An alternative to the classification-based methods, are the ones that use a probabilistic formulation to model the co-occurrence of image features and labels [21, 25, 34, 61, 65, 93, 106, 111, 132, 153, 185, 194]. The general idea is to use a Bayesian framework to estimate the posterior distribution of each of the possible labels, given the observation of features from the image, defined as $p(w_m|\mathbf{r})$, where w_m is the m-th possible annotation and **r** is the set of image features. Then, the estimated probabilities are ranked and the L labels with the highest posterior are selected to annotate the image. This makes it possible to assign multiple labels to the same image.

The main difference between methods is the way they define and estimate the joint probability of words and images. Preliminary works on this topic treated the annotation problem as a translation one, where the goal was to translate from image features to keywords [21, 61, 132, 153]. The framework of these models is very simple. During the training phase, the images are divided into regions and these regions are clustered in order to obtain a set of representative centroids. Then, the regions are associated with the closest centroid and each of them inherits the labels of the image it came from. The final step consists of either computing the co-occurrence between labels and centroids [132], or to estimate the probability of the label w_m given a centroid [61].

The main downside of the previous methods is that it is not appropriate for a region to inherit all of the labels associated with the entire image. This problem was addressed in [65, 93, 106], where instead of computing the joint probability of a label and a region, these methods estimate the joint probability of a label and the entire image. In order to accomplish this, it is necessary to learn a model that relates labels and images and another one that relates images and regions features. The cross media relevance model (CMRM) proposed by Jeon et al [93] treats this problem as a discrete one. Similarly to the translation problems described above, they decompose the images into regions and then cluster them, in order to obtain a set of centroids. As before, the image regions are associated with the closest centroid. The models for both labels and regions are computed using interpolation. Lavrenko et al. [106] extended the previous model to a continuous formulation, proposing the continuous relevance model (CRM). Feng et al. [65] propose a continuous way of addressing the computation of the models, using the multiple-Bernoulli distribution to model the joint distribution of labels and images. This model is called multiple-Bernoulli relevance model (MBRM).

Another type of probabilistic methods are those that define the relationship between region features and text labels in an indirect way, using hidden variables. An example is the correspondence latent Dirichlet allocation (corr-LDA) algorithm [25], which belongs to the family of generative methods. This method assumes that there is a set of hidden variables called *topics* that are simultaneously associated with a distribution over region features and possible text labels. Under this assumption, it is possible to estimate a joint distribution of text labels and features and, consequently, the desired labeling probabilities $p(w_m|\mathbf{r})$.

Similarly to classification methods, probabilistic ones also have some limitations. Translation methods and the CMRM both require a first clustering step, which can make their performance very sensitive to clustering errors [65]. MBRM and CRM do not require a preliminary clustering step. However, these methods cannot be applied when one is trying to simultaneously label images and regions. corr-LDA solves this issue, but requires a priori definition of the number of topics and the selection of a suitable probabilistic formulation to model the region features. All of the described methods require a preliminary segmentation of the image into regions, which makes them sensitive to noisy regions.

9.3 Problem Formulation

Taking into account the considerations made by a committee of dermatologists in [59], it is possible to define two main requirements that the proposed system must fulfill (see Figure 9.1):

i) Provide relevant clinical information to the dermatologists, *i.e.*, replicate their identification of dermoscopic criteria. The system should provide a set of text labels stating which are the clinical criteria that are present in the lesion and associate those labels with specific regions (region annotation), such that they can be checked by the physicians.

ii) Diagnose the lesions, basing that decision on the detected clinical criteria. This ensures that the features used by the system have a medical meaning, making it possible for the dermatologist to understand and validate the automated diagnosis.

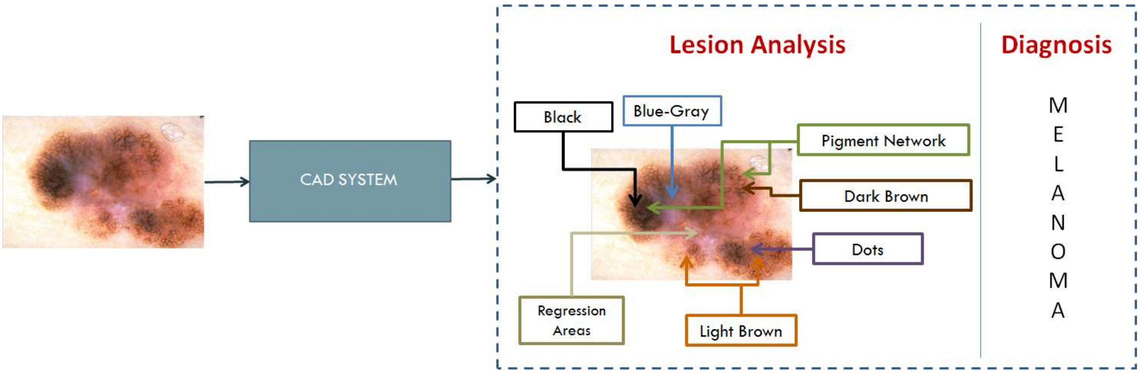
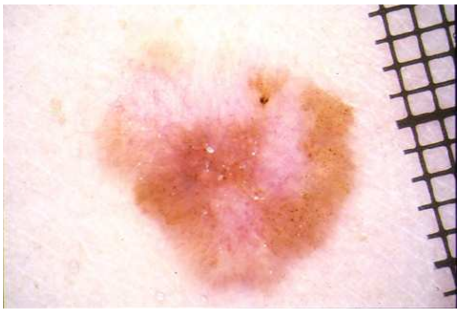


Figure 9.1: Desired output of a clinically inspired system.

The aforementioned requirements raise a set of problems. First, the medical criteria involve the detection of very subtle structures. Second, the development of strategies to detect clinical criteria usually requires large datasets of images with detailed information: text annotations stating which are the criteria that can be found in the lesion and corresponding region segmentations. The segmentations are used in the training of several algorithms, as mentioned in Section 9.2.1. Unfortunately, most datasets lack segmentations and only provide text labels, as exemplified in Figure 9.2, since performing the latter is seen as a time consuming and subjective task by the experts. Finally, it is not easy for a computer to convert the detected criteria into information that can be used to automatically diagnose melanomas.

Each of the aforementioned problems is addressed in this chapter. The lack of reliable segmentations associated to the clinical features is tackled using an image annotation approach. Section 9.2.2 presented several annotation algorithms, each with different properties. All of the presented approaches have pros and cons. Supervised classification strategies can range from simple binary classification models to more complex methods, where statistical relations between the different labels are considered for the annotation process. Region annotation is performed using methods based on the MIML framework, which has been shown to achieve interesting results. However, MIML methods have a series of drawbacks, such as the need to define the number of regions per image or the

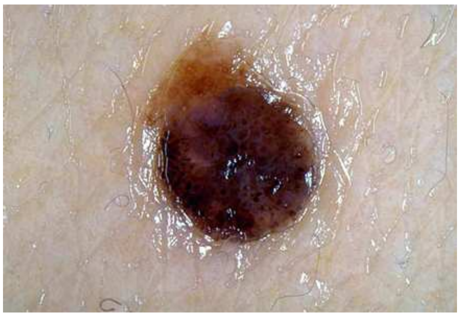


Colors: Dark brown, light brown, red, and white.

Texture Structures: Dots.

Color Structures: Regression areas.

Diagnosis: Melanoma.



Colors: Dark brown, black,light brown, and blue.

Texture Structures: Pigment network and dots.

Color Structures: Blue-whitish veil.

Diagnosis: Benign.

Figure 9.2: Images and annotations provided by dermatologists [9].

inclusion of a preliminary clustering step, and the decomposition of the learning process into single class learning procedures. Some of the probability based methods are also suitable for region labeling (*e.g.*, [25, 61, 132]). The co-occurrence method [132] and the translation method [61] suffer from assuming that each of the regions inherits all of the labels of their mother image. corr-LDA [25] does not make this assumption, thus it is an appropriate choice for this work.

Since some medical criteria can be difficult to detect, it is important to select a subset that play an important role in the diagnosis. The selected criteria correspond to different characteristics of the lesions and can be divided into three classes (as exemplified in Figure 9.2):

- The six colors (C) - dark and light browns, blue-gray, black, white, and red [182];
- Two texture structures (TS) - pigment network and dots/globules [182];
- Two color structures (CS) - blue-whitish veil and white regression areas [7].

Finally, all of these criteria are used to extract appropriate features for lesion diagnosis.

There are two main differences between the proposed system and other clinically inspired methods: i) most of the methods focus on the detection of one or two clinical criteria [103], while the proposed method detects a larger number of criteria that characterize different aspects of the lesion; and ii) few methods try to diagnose melanomas using the clinical features [103], which is performed in this chapter.

9.4 System Overview

This section succinctly describes the clinically inspired CAD system proposed in this chapter [14]. The sequential framework of the system is similar to the analysis performed by dermatologists, *i.e.*, first the system tries to identify the presence or relevant dermoscopic criteria and then performs a diagnosis using this information. Figure 9.3 exemplifies the pipeline of the system.

The first step consists of dividing the image into smaller regions (see Figure 9.3), each of them characterized by a feature vector $\mathbf{r}_n$. An image is assumed to be represented by a set $\mathbf{r} = \{\mathbf{r}_1, ..., \mathbf{r}_N\}$,

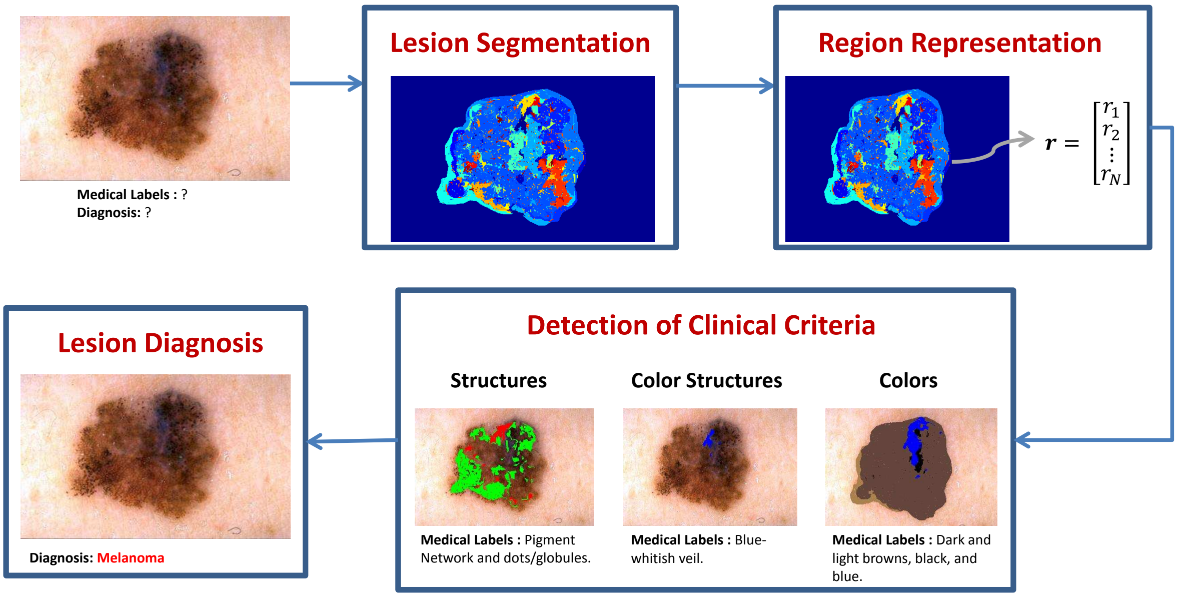

Medical Labels : ?
Diagnosis: ?
Lesion Segmentation
Region Representation
$r = \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_N \end{bmatrix}$
Detection of Clinical Criteria
Structures
Color Structures
Colors
Medical Labels : Pigment Network and dots/globules.
Medical Labels : Blue-whitish veil.
Medical Labels : Dark and light browns, black, and blue.
Lesion Diagnosis
Diagnosis: Melanoma

Figure 9.3: Clinically inspired CAD system.

which comprises the feature vectors from all of the N regions. The segmentation strategy will be discussed in Section 9.5. The features that characterized the regions are discussed in Section 9.6. It is assumed that each of the images has one or more text labels.

The next step is the identification of the dermoscopic criteria that are present in the lesion. This is a two fold task, as exemplified in Figure 9.3, where the medical criteria are associated with one or more regions (local labels) and text labels are produced for the entire image (global labels) . The detection of the criteria is based on the generative model corr-LDA [25], which is estimated using a database of weakly annotated images. The probabilistic formulation of corr-LDA as well as specific aspects of the annotation process will be discussed in Section 9.7.

The final block of the system classifies the lesion as melanoma or benign using information extracted from the detected medical criteria. This task requires the use of a classification algorithm, which is learned using a dataset of dermoscopy images diagnosed by experts. The learning process of the classifier works as follows. First, the previously estimated corr-LDA models are applied to the training images. Then, new features are extracted from their output. Finally, these features are used to train the classifier. In the case of new images, their corr-LDA outputs are used as features and then the classifier is applied to predict the diagnosis. Detailed information about the classification approach is provided in Section 9.8.

The main advantage of the proposed system is its ability to interact with the dermatologist, since the automated diagnosis relies on descriptors that are medically inspired and can be checked by clinicians. This allows them to understand and validate the suggested lesion diagnosis. Furthermore, its sequential framework is similar to the analysis performed by an expert: first look for several dermoscopic criteria and then diagnose the lesion. These two characteristics of the proposed system make it valuable for the medical community and make it significantly different from other systems found in literature [59].

9.5 Image Segmentation

This section addresses the strategy used to divide the skin lesion into smaller homogeneous regions and the type of features used to describe those regions.

9.5.1 Overview of Image Segmentation Methods

Different approaches have been used to perform image segmentation in the context of image annotation problems. These approaches range from a simple uniform grid method, where the image is divided into small square blocks, to more complex strategies that try to split the image into disjoint regions that preserve some kind of homogeneity [205].

Applying a regular grid to split the lesions into patches of a predefined size is the simplest and less computationally expensive segmentation strategy. However, this method leads to misleading segmentations, since it does not separate well different color and texture components of an image. In the case of dermoscopy, a single patch may contain more than one color or structure. This may lead to the extraction of inaccurate features and, consequently, poor criteria detections.

It is possible to deal with the aforementioned problem by using more elaborate segmentation algorithms, where the goal is to obtain regions that are homogeneous regarding color and/or texture properties. These algorithms can be divided into different groups: clustering algorithms, statistical models, graph-based algorithms, and region growing methods [205]. All of these approaches share a preliminary feature extraction step, where each pixel in the image is characterized by a feature vector. These vectors can comprise information about color, texture or both. The next step consists of grouping the pixels into regions, which can be done in different ways. Clustering (*e.g.*, SLIC [4]) and region growing methods (*e.g.*, JSEG [55]) divide the pixels into a predefined number of regions (clusters). Statistical models are based on the estimation of probability density functions over the pixel's features. Examples of this kind of methods are the blobworld [35], mean-shift [46], and quick-shift [190] algorithms. Graph-based methods, such as [64,170], define the pixels as vertices of a graph and a measure of similarity between feature vectors of neighbor pixels as their edge weights. Based on this formulation, the segmentation problem becomes one of graph partitioning, where the goal is to separate the vertices/pixels into disjoint sets such that the similarity between them is minimized. Each of the aforementioned methods lead to regions that differ in terms of shape, size, and characteristics.

9.5.2 Segmentation Approach

The goal of this chapter is to detect criteria that have specific colors and/or texture properties, meaning that the segmentation algorithm must provide regions that are homogeneous regarding these two properties. Moreover, structures like pigment network may appear at different scales. Hence, it is also important to ensure that the obtained regions are scale invariant. Finally, there is no constraint regarding the shape, the size, and the number of the regions, since the distribution of the criteria inside the lesion is unknown.

The method proposed by Felzenszwalb and Huttenlocher [64] is adopted in this chapter to fulfill the aforementioned requirements. This is a graph-based algorithm that assumes that each of the pixels in the image is one vertice of a graph, and that neighbor pixels are linked through edges. The weight of each edge is given by the Euclidean distance between the feature vectors of the two neighbor pixels. Connected vertices/pixels are combined in the same region if they are similar, *i.e.*, if their corresponding edge weight is lower than a given threshold δ. The value of δ defines the size of the obtained regions and is influenced by the resolution of the images. It was experimentally found that setting $\delta = \frac{(R^d + C^d)}{130}$, where $R^d \times C^d$ is the resolution (size) of image d, led to a good trade off between the size of the regions and the computation time.

Each of the pixels in the image is characterized by a feature vector that describes its color and texture content. These features are extracted using a strategy that enforces scale invariance, where the main idea is to determine the optimal scale for each pixel and then extract its features at that scale [35]. This optimal scale is estimated using an image property called polarity, which measures the extent to which all of the gradient vectors in the neighborhood of a pixel point in the same direction. The polarity value of a pixel is computed with respect to the most common orientation ϕ of the gradients in its neighborhood, as follows

$$p_s = \frac{|E_+ - E_-|}{E_+ + E_-}, \tag{9.1}$$

where

$$\begin{aligned} E_+ &= \textstyle\sum_{x,y} G_s(x,y)[\nabla I.n]_+ \\ E_- &= \textstyle\sum_{x,y} G_s(x,y)[\nabla I.n]_- . \end{aligned} \tag{9.2}$$

$G_s(x,y)$ is a Gaussian smoothing kernel with variance (scale) $\sigma^2 = s^2$, ∇I is the image gradient computed using the first difference approximation in each direction, and n is a unit vector perpendicular to the preferential orientation ϕ. The terms $[.]_+$ and $[.]_-$ are the rectified positive and negative parts of the argument. The polarity at every pixel in the image is computed for $s_k(x,y) = k/2$, $k = 0, 1, ..., 7$. The best scale $s^*(x,y)$ of each pixel is selected as the first value of $s_k(x,y)$ for which the difference between consecutive polarity values is less than 2% [35].

Given the best scales at each pixel, it is possible to compute the features as follows. The color information of a pixel are its L*a*b* components. These components are determined after performing spatial averaging using a Gaussian smoothing kernel with $\sigma^2 = s^*(x,y)^2$. The texture information is characterized by the contrast, contrast × polarity, and anisotropy × contrast [35]. Contrast and anisotropy of a pixel (x,y) are computed using the second moment matrix

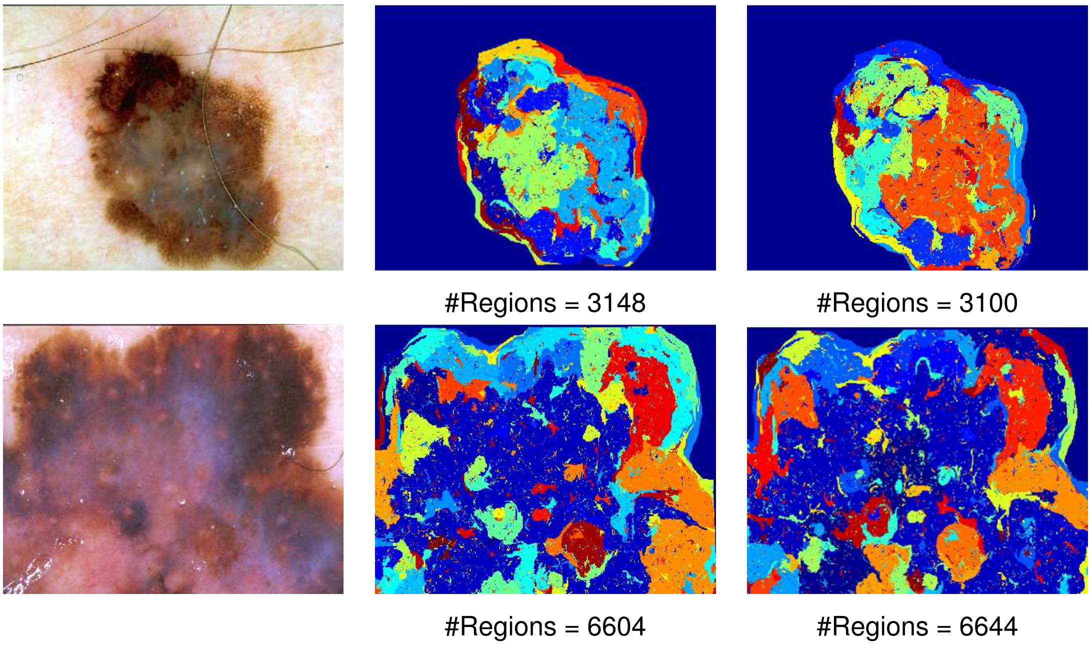
$$M_s(x,y) = G_s(x,y) * (\nabla I)(\nabla I)^T, \tag{9.3}$$

where $s = s^*(x,y)$ is the pixel's best scale. From this matrix, anisotropy and contrast at each pixel (x,y) are respectively defined as:

$$a(x,y) = 1 - \tfrac{\lambda_2}{\lambda_1} \quad , \quad c(x,y) = 2\sqrt{\lambda_1 + \lambda_2}, \tag{9.4}$$

where λ_1, λ_2 are the eigenvalues of $M_s(x,y)$.

Two segmentations are performed for each image, one using only color features and the other using color and texture features, as exemplified in Figure 9.4. The first segmentation is used to train/test the corr-LDA associated with the color criteria, while the remaining are used to train the models associated with color and texture structures.


#Regions = 3148

#Regions = 3100

#Regions = 6604

#Regions = 6644

Figure 9.4: Image segmentation: original image (left), segmentation using color features (mid), and segmentation using color and texture features (right). Each color label represents a different region.

9.6 Region Representation

After segmenting the lesions, as exemplified in Figure 9.4, each of the $1, \ldots, N$ regions is characterized by a feature vector $r_n \in \mathbb{R}^f$. This feature vector comprises information about color, texture or both, depending on the type of dermoscopic criteria. Therefore, an image d is represented by a set $\mathbf{r}^d = \{r_1^d, ..., r_N^d\} \in \mathbb{R}^{f \times N^d}$ of N^d vectors (recall from Figure 9.4 that each lesion is segmented into a different number of regions). Since the best type of features to describe the regions is unknown, combinations of the following descriptors were tested for each type of criterion.

- **Colors:** The mean color vector in the HSV space (μ_{HSV}).
- **Texture structures:** The regions are described using texture features. In this chapter the tested features are the mean contrast (μ_c) and mean contrast $\times$ anisotropy (μ_{ca}), the mean (μ_g) and standard deviation (σ_g) of the gray level values in the region, and statistics computed using the directional filters described in Chapter 3. These filters are computed at different orientations $\theta_i \in [0, \pi]$, $i = 0, ..., 9$, with the impulse response for direction θ_i given by

$$h_{\theta_i}(x, y) = G_1(x, y) - G_2(x, y), \tag{9.5}$$

where G_k is a Gaussian filter:

$$G_k(x,y) = C_k \exp\{-\frac{x'^2}{2\sigma_{x_k}^2} - \frac{y'^2}{2\sigma_{y_k}^2}\}, k = 1, 2, \tag{9.6}$$

where $\sigma_{x_1} = 40$, $\sigma_{y_1} = 40$, $\sigma_{x_2} = 3$, $\sigma_{y_2} = 0.5$, and In (9.6) C_k is a normalization constant and the values of (x', y') are related with (x, y) by a rotation of amplitude θ_i.

$$\begin{aligned} x' &= x\cos\theta_i + y\sin\theta_i\,, \\ y' &= y\cos\theta_i - x\sin\theta_i\,. \end{aligned} \tag{9.7}$$

The size of the masks are 11×11. The output of the directional filters (9.5) is computed for all the directions, and the maximum and minimum values are kepts for each pixel. The regions are described by the mean and standard deviation of these values (μ_M, σ_M, μ_m, and σ_m).

- **Color Structures:** These structures simultaneously exhibit color and texture properties. The color of the regions is characterized using μ_{HSV}, while the texture is characterized using the features (μ_c, μ_{ca}) and (μ_M, σ_M, μ_m, σ_m).

All the clinical criteria and the tested features/descriptors are summarized in Table 9.1. After computing the features for all the N^d regions, it is possible to describe image d as a set **r** $= \{r_1^d, ..., r_{N^d}^d\}$ of N^d of feature vectors.

Table 9.1: Feature summary - for details see Section 9.6.

	HSV	**Contrast, Anisotropy**	**Gray-Level Image**	**Directional Filters**
Criteria	μ_{HSV}	μ_c, μ_{ca}	μ_g, σ_g	μ_M, σ_M, μ_m, σ_m
Color	✓			
Texture Structures		✓	✓	✓
Color Structures	✓	✓		✓

9.7 Detection of Clinical Criteria

This section gives and overview of the probabilistic formulation of corr-LDA [25], the parameter estimation process, and the methodology used to associated the regions and the image with the clinical criteria.

9.7.1 Annotation Model - Correspondence Latent Dirichlet Allocation

The goal of image annotation methods is to find a relationship between text labels and image features, such that a computer can replicate the annotation process in new images. This is similar to the goal of this chapter, since the objective is to assign clinical labels to dermoscopic images. corr-LDA is a generative annotation model that assumes the existence of a sequence of local observations (region features) $r_1, ..., r_N$ and global text label w, associated with an image. For the sake of simplicity lets assume that w is a single label belonging to the set $\{w_1, ..., w_M\}$. The relationship between the local observations and the global label is based on the existence of N latent variables (called *topics*) $z_1, ..., z_N$, each one associated with a region. These topics are the core of the corr-LDA, as will be seen in the sequel.

It is assumed that the observation r_n depends on the topic z_n, as is depicted in Figure 9.5. Each r_n is generated by an observation model $p(r_n|z_n, \Omega_{z_n})$, with hyperparameter(s) Ω. The main difficulty concerns the generation of the global label w, since the regions associated with it are unknown. corr-LDA deals with this problem by randomly selecting a region $n \in \{1, 2, .., N\}$, and generating the label according to a probabilistic model conditioned on the topic of the region $p(w|z_n)$ (see Figure 9.5). This formulation makes it possible to establish a relationship between the local observations and a global text label provided for the entire image.

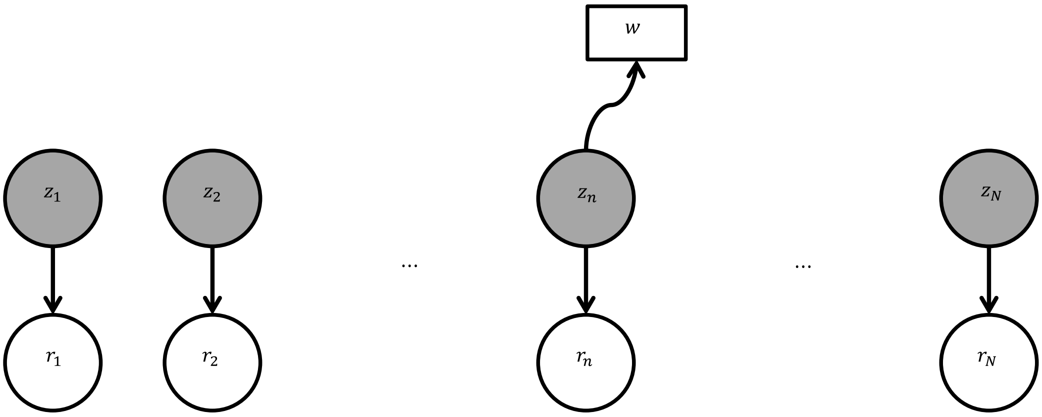


Figure 9.5: Generative process of corr-LDA.

The probabilistic formulation of corr-LDA is the following: i) the topics are randomly generated by a multinomial distribution $z_n \sim$ Mult(θ) with parameter $\theta \in \mathcal{R}^K$, obtained from a Dirichlet distribution of hyperparameter $\alpha \in \mathcal{R}^K$, where K is a value defined by the user that corresponds to the number of different topics that can be considered by the model; ii) the observations r_n are generated by an observation model with hyperparameter(s) Ω_{z_n}; and iii) the global label w is generated by a multinomial distribution $w \sim$ Mult(β_{z_n}).

The rest of this section is inspired in [25, 26], except what concerns the Von Mises distribution. The complete generative process from an image d, now assuming more than one global label, is summarized below [25]:

1. For an image d, sample a topic distribution $\theta \sim$ Dirichlet(α).
2. For each of the N^d image regions:
   (a) Sample a topic $z_n \sim$ Mult(θ).
   (b) Sample a region descriptor $r_n \sim p(r|z_n, \Omega)$ from a distribution conditioned on z_n.
3. For each of the M^d global labels w_m:
   (a) Sample a region indexing variable $y_m \sim$ Unif$(1, ..., N)$.

(b) Sample an annotation $w_m \sim p(w|y_m, \mathbf{z}, \beta)$ from a multinomial distribution conditioned on the z_{y_m} topic.

This generative process leads to the following joint distribution of region features, labels, and hidden variables

$$p(\mathbf{r}, \mathbf{w}, \theta, \mathbf{z}, \mathbf{y}|\alpha, \beta, \Omega) = p(\theta|\alpha).\left(\prod_{n=1}^{N} p(z_n|\theta)p(r_n|z_n, \Omega)\right).\left(\prod_{m=1}^{M} p(y_m|N)p(w_m|y_m, \mathbf{z}, \beta)\right). \qquad (9.8)$$

During the training phase it is necessary to estimate all of the model parameters α, $\Omega = \{\Omega_1, ..., \Omega_K\}$, and $\beta = \{\beta_1, ..., \beta_K\}$, using a training set of D weakly annotated images. The traditional approach consists of using a maximum likelihood (ML) formulation, which requires the computation of

$$p(\mathbf{r}, \mathbf{w}|\alpha, \beta, \Omega) = \prod_{d=1}^{D} \int_{\theta^d} \sum_{\mathbf{z}^d} \sum_{\mathbf{y}^d} p(\mathbf{r}^d, \mathbf{w}^d, \theta^d, \mathbf{z}^d, \mathbf{y}^d|\alpha, \beta, \Omega). \qquad (9.9)$$

This expression cannot be analytically computed. Blei and Jordan [26] address this issue using a variational method to estimate the parameters. This approach starts by introducing a new set of independent variational parameters, each associated with a distribution over a specific hidden variable of the original model. The variational parameters are image specific, *i.e.*, each image will be associated with a different set of variational parameters. These parameters, here identified as (γ, ϕ, λ), allow the definition of a factorized distribution of the hidden variables $(\theta, \mathbf{z}, \mathbf{y})$

$$q(\theta, \mathbf{z}, \mathbf{y}) = q(\theta|\gamma).\left(\prod_{n=1}^{N} q(z_n|\phi_n)\right).\left(\prod_{m=1}^{M} q(y_m|\lambda_m)\right). \qquad (9.10)$$

The factorized distribution can be introduced in the original log-likelihood using Jensen's inequality:

$$\begin{aligned}\sum_{d=1}^{D} \log p(\mathbf{r}^d, \mathbf{w}^d|\alpha, \beta, \Omega) &= \sum_{d=1}^{D} \log \int_{\theta^d} \sum_{\mathbf{z}^d} \sum_{\mathbf{y}^d} p(\mathbf{r}^d, \mathbf{w}^d, \theta^d, \mathbf{z}^d, \mathbf{y}^d|\alpha, \beta, \Omega)\mathrm{d}\theta \\ &= \sum_{d=1}^{D} \log \int_{\theta^d} \sum_{\mathbf{z}^d} \sum_{\mathbf{y}^d} \frac{p(\mathbf{r}^d, \mathbf{w}^d, \theta^d, \mathbf{z}^d, \mathbf{y}^d|\alpha, \beta, \Omega)q(\theta^d, \mathbf{z}^d, \mathbf{y}^d)}{q(\theta^d, \mathbf{z}^d, \mathbf{y}^d)}\mathrm{d}\theta \\ &\geq \sum_{d=1}^{D} \mathbb{E}_q[\log p(\mathbf{r}^d, \mathbf{w}^d, \theta^d, \mathbf{z}^d, \mathbf{y}^d|\alpha, \beta, \Omega)] - \mathbb{E}_q[\log q(\theta^d, \mathbf{z}^d, \mathbf{y}^d)], \qquad (9.11)\end{aligned}$$

where $\mathbb{E}_q$ is the expected value according to the variational distribution $q(\theta, \mathbf{z}, \mathbf{y})$. The right side of the equation gives an overall lower bound of the log-likelihood $\mathcal{L}(\mathbb{D})$, for a set of D images. It is possible to observe that this bound is the sum of the individual lower bounds of each image, $\mathcal{L}(\gamma^d, \phi^d, \lambda^d; \alpha, \beta, \Omega)$, which can be decomposed as follows:

$$\begin{aligned}\mathcal{L}(\gamma^d, \phi^d, \lambda^d; \alpha, \beta, \Omega) &= \mathbb{E}_q[\log p(\theta^d|\alpha)] + \mathbb{E}_q[\log p(\mathbf{z}^d|\theta^d)] \\ &+ \mathbb{E}_q[\log p(\mathbf{r}^d|\mathbf{z}^d, \Omega)] + \mathbb{E}_q[\log p(\mathbf{y}^d|N^d)] \\ &+ \mathbb{E}_q[\log p(\mathbf{w}^d|\mathbf{y}^d, \mathbf{z}^d, \beta)] - \mathbb{E}_q[\log q(\theta^d|\gamma^d)] \\ &- \mathbb{E}_q[\log q(\mathbf{z}^d|\phi^d)] - \mathbb{E}_q[\log q(\mathbf{y}^d|\lambda^d)]. \qquad (9.12)\end{aligned}$$

Each of the terms in (9.12) can be expanded into explicit functions of the model (α, β, Ω) and variational $(\gamma^d, \phi^d, \lambda^d)$ parameters. For the sake of simplicity, the expression of each of the terms of $\mathcal{L}(\gamma^d, \phi^d, \lambda^d; \alpha, \beta, \Omega)$ is included in Appendix C.

In the end, the problem is transformed into one of finding the best set of parameters that maximizes $\mathcal{L}$, and can be solved using a variational Expectation-Maximization (EM) algorithm [98]:

- **E-step:** Estimate the specific variational parameters of each image d in the training set. The update equations of the parameters are obtained by taking derivatives of $\mathcal{L}(\gamma^d, \phi^d, \lambda^d; \alpha, \beta, \Omega)$ with respect to each of the parameters $(\gamma^d, \phi^d, \lambda^d)$ and setting them to zero.
- **M-Step:** Estimate the model parameters (common to all training images) by maximizing the overall lower bound $\mathcal{L}(\mathbb{D})$, with respect to (α, β, Ω).

These two steps are performed until $\mathcal{L}(\mathbb{D})$ converges. The updates equations of the variational $(\gamma^d, \phi^d, \lambda^d)$ and model (α, β) parameters are the same ones as proposed in [25, 26]. The update equations for Ω depend on the distributions $p(r_n|z_n, \Omega_{z_n})$. In [25] these distributions are defined as multivariate Gaussian. However, this kind of distribution is not suitable to model all types of features. An example are the features that comprise periodic angular information, such as the Hue channel of the HSV color space [32]. Therefore, two distributions are applied in this work, according to the features r_n used to describe the regions. If μ_{HSV} is included in the feature vector, then $p(r|z_n, \Omega_{z_n})$ is a von-Mises multivariate Gaussian distribution [32]

$$p(r_n|z_n, \Omega_{z_n}) = \nu(\mathsf{H}_n|z_n, \tau_{z_n}, \varepsilon_{z_n}) G(r'_n|z_n, \mu_{z_n}, \Sigma_{z_n}), \tag{9.13}$$

where G is a multivariate Gaussian, and r' defined the feature vector of the region, without the feature corresponding to the H channel. ν is a von-Mises distribution

$$\nu(\mathsf{H}_n|z_n, \tau_{z_n}, \varepsilon_{z_n}) = \frac{1}{2\pi I_0(\varepsilon_{z_n})} e^{\varepsilon_{z_n} \cos(\mathsf{H}_n - \tau_{z_n})}, \tag{9.14}$$

where the normalization factor I_0 is the modified zero-order Bessel function of the first kind and $\varepsilon_{z_n} \geq 0$ denotes the concentration of the distribution around the mean τ_{z_n}. In this case, Ω_{z_n} comprises four parameters $(\mu_{z_n}, \Sigma_{z_n}, \tau_{z_n}, \varepsilon_{z_n})$ that must be estimated. Otherwise, $p(r_n|z_n, \Omega_{z_n})$ is a Gaussian distribution. For completeness, all update equations can be found in Appendix C.

9.7.2 Region and Image Labeling

This section addresses local and global labeling using the estimated model. To annotate the regions (local labeling) of a new image, *i.e.*, to associate them with a medical criteria, it is necessary to compute the following probability for each of the admissible labels w

$$p(w|r_n) \propto \sum_{z_k} q(z_k|\phi_{nk}) p(w|z_k, \beta), \tag{9.15}$$

where ϕ_{nk} is the variational parameter of region n associated with topic k, and $q(z_k|\phi_{nk})$ is a multinomial distribution. The label w with the highest probability $p(w|r_n)$ is then selected [25]. This is clearly a greedy formulation, where all the labels compete to annotate a region and only one is selected. However, this assumption is not valid in the case of dermoscopy, where more than one label can be associated with the same region. An example is Figure 9.1: the color labels *dark brown* and *black* share regions with the texture label *pigment network*. This issue is addressed in this thesis using two different strategies:

i) Train three corr-LDA models, one for each class of medical criteria (colors, texture structures, and color structures).

ii) Train two corr-LDA model, one for color and another for all of the structures.

The traditional approach used in corr-LDA to obtain global image labels consists of computing the following image label probability [25]

$$p(w|\mathbf{r}) \propto \sum_{n=1}^{N} \sum_{z_k} q(z_k|\phi_n) p(w|z_k, \beta), \tag{9.16}$$

for all the candidate labels. Then, the labels are ranked from the most to the less probable, and a small predefined number of them is selected. This approach restricts the number of labels per image, making it unsuitable for the annotation of dermoscopy images, where any number of criteria may be present. A clear example are the color labels, where it is possible to find just one color or all six of them in a single lesion. To tackle this issue, the image labeling process is reformulated as a set of classification problems.

In the case of the colors model, the image receives a color label if the following area ratio is above an estimated threshold

$$\delta_c = \frac{A^c_{regions}}{A_{lesion}}, \tag{9.17}$$

where $A^c_{regions}$ is the total area of the regions annotated with color c and A_{lesion} is the area of the lesion. This is the same criterion adopted in Chapter 8 for the detection of colors.

In the case of texture and color structures, the idea is to use a set of classification algorithms to predict the label, each of them trained to predict one of the possible labels. The features used by the classifiers are the outputs of corr-LDA: the image label probability (9.16), computed for all the labels, and the average number of regions per topic $\eta \in \mathbb{R}^K$, where each position k is given by

$$\eta_k = \alpha_k - \gamma_k. \tag{9.18}$$

The latter descriptor is based on the assumption that the k-th position of variational parameter γ corresponds approximately to the k-th position of model parameter α plus the expected number of patch features that were generated by the k-th topic [26]. Two classification algorithms are tested in this chapter: random forests and SVM.

9.8 Lesion Diagnosis

The previous section described the strategy used to obtain a medical description of the lesions, both in the form of global text labels, which describe the entire lesion, and of local labels associated to specific regions. This section provides an insight of how to use the detected information to discriminate between benign and malignant lesions.

In order to be able to diagnose skin lesions it is necessary to answer to two questions:

i) How to convert the annotations into an appropriate descriptor that can be used by machine learning algorithms?

ii) How to combine the information of the different corr-LDA models in order to obtain a final diagnosis?

The annotations obtained using corr-LDA as well as other outputs of this algorithm are used to compute a lesion descriptor that comprises the following information:

- **Present/Absent criteria:** consists of three binary vectors $\mathbf{f}_C \in \mathbb{R}^6$, $\mathbf{f}_{TS} \in \mathbb{R}^2$, and $\mathbf{f}_{CS} \in \mathbb{R}^2$, such that the i-th position of each of these vectors is equal to 1 if the image was annotated with the i-th label, and 0 otherwise.

- **Label distribution:** consist of the conditional probabilities $p(w|\mathbf{r})$ (9.16), which provide an estimate of the labels probabilities in a given lesion.

- **Average number of regions per topics:** comprises the vector $\eta \in \mathcal{R}^K$ (9.18), for each of the trained corr-LDA models.

More than one corr-LDA model is used in this chapter to obtain the annotations, as discussed in Section 9.7.2. Thus, it will be necessary to combine the information of the different models. The easiest strategy would be to compute the aforementioned features for each of the models and then combine all the information into a single descriptor (early fusion). However it has been shown that early fusion is not appropriate to combine different types of features [104], such as those extracted from the set of corr-LDA models. In Chapter 7 the classification results showed that late fusion [104] outperformed early fusion. Thus, this strategy will also be applied in this chapter.

As a reminder, late fusion consists of combining the outputs of different classifiers, in order to obtain a final decision. Here, each classifier will be trained using information of one of the corr-LDA models. The outputs of the different classifiers are a set of scores in the interval $[0, 1]$, where the lesion is considered malignant if its score is above 0.5. These scores are the input of a final classifier that predicts the final diagnosis.

Two classification algorithms are tested as candidates for the first line of classifiers: SVM with RBF kernel and random forests. The scores of the different classifiers are combined using two different strategies: logistic regression (LR) and median rule (MR) [100].

9.9 Experimental Results

9.9.1 Dataset and Evaluation Metrics

The experiments were performed using a dataset of 804 melanocytic lesions (241 melanomas) extracted from EDRA [9]. This was the only dataset used in this chapter because it contained much more training examples than PH^2. Nonetheless, it was still necessary to increase the number of images when compared with the previous chapters, in order to include sufficient examples of each type of dermoscopic criteria. All of the images were analyzed by several experts during a consensus meeting. Each image is associated with: i) a set of text labels stating which are the observed criteria; and ii) a malignant/ benign diagnosis. Text labels associated with texture and color structures are available for all the images. However, color labels are only available for a subset of 344 images.

All of the images were pre-processed in order to remove acquisition artifacts and skin hair using the algorithm developed in Chapter 3, and their colors were normalized as proposed in Chapter 6. Manual segmentations were used to separate the lesions from healthy skin.

The detection of the medical criteria was evaluated using three metrics: precision (Pre), recall (Re), and the $F1$ score. These metrics are used to compare the global labels provided by the automatic system against those of the experts. Lesion diagnosis is evaluated using sensitivity (SE), specificity (SP), and the cost index (S) (4.19). All of the metrics were computed using nested 10-fold cross validation.

9.9.2 Detection of Medical Criteria

Different experiments were conducted to optimize this block. The goal of the experiments was:

i) Assess which is the best subset of region features (recall Section 9.6).

ii) Select the best classifier to obtain the global labels (SVM or random forests).

iii) Compare the performance of training a corr-LDA model for texture structures and another for color structures against a model that comprises all the structures.

The number of topics of the corr-LDA models was tuned in the set $K \in \{40, 50, ..., 300\}$. For each of the trained corr-LDA models was then trained a set of classifiers to predict the global text labels. The hyperparameters of each classifier were optimized as follows. In the case of random forests, the number of trees was searched in the set $T \in \{1, 2, ..., 50\}$, while in the case of SVM two hyperparameters were optimized: the width of the RBF kernel $\rho \in \{2^{-12}, 2^{-5}, ..., 2^{12}\}$ and the penalty term $C \in \{2^{-6}, 2^{-4}, ..., 2^{6}\}$ given to the soft margin. All of optimal hyperparameters were selected by nested cross validation. This resulted in a total of 486000 configurations for random forests and 1579500 configurations for SVM. In all of the experiments the features were normalized, as described in Chapter 4.

Table 9.2 shows the performance of the best annotation models, as well as the best configurations. Results for the remaining configurations can be found in Appendix D. The first three rows show the results obtained after training a model for each type of criteria. Most of the colors were detected with good scores. However, the performance scores for the red and white colors are lower than for the others, as expected, since there are few examples of these colors on the dataset. Regarding the texture structures' model (2nd row), the best results were obtained when all of the features are used (μ_g, σ_g, μ_c, μ_{ca}, μ_M, σ_M, μ_m, σ_m). This leads to Re scores above 80% for both of the criteria using Random Forests. The color structures' model (3rd row) is the one that achieved the worse scores. This was expected as the number of examples of each of the color structures is smaller than

Table 9.2: Detection results and best configurations for the four annotation models - * identifies the results of the model that combines color and texture structures. In **bold** we highlight the best results.

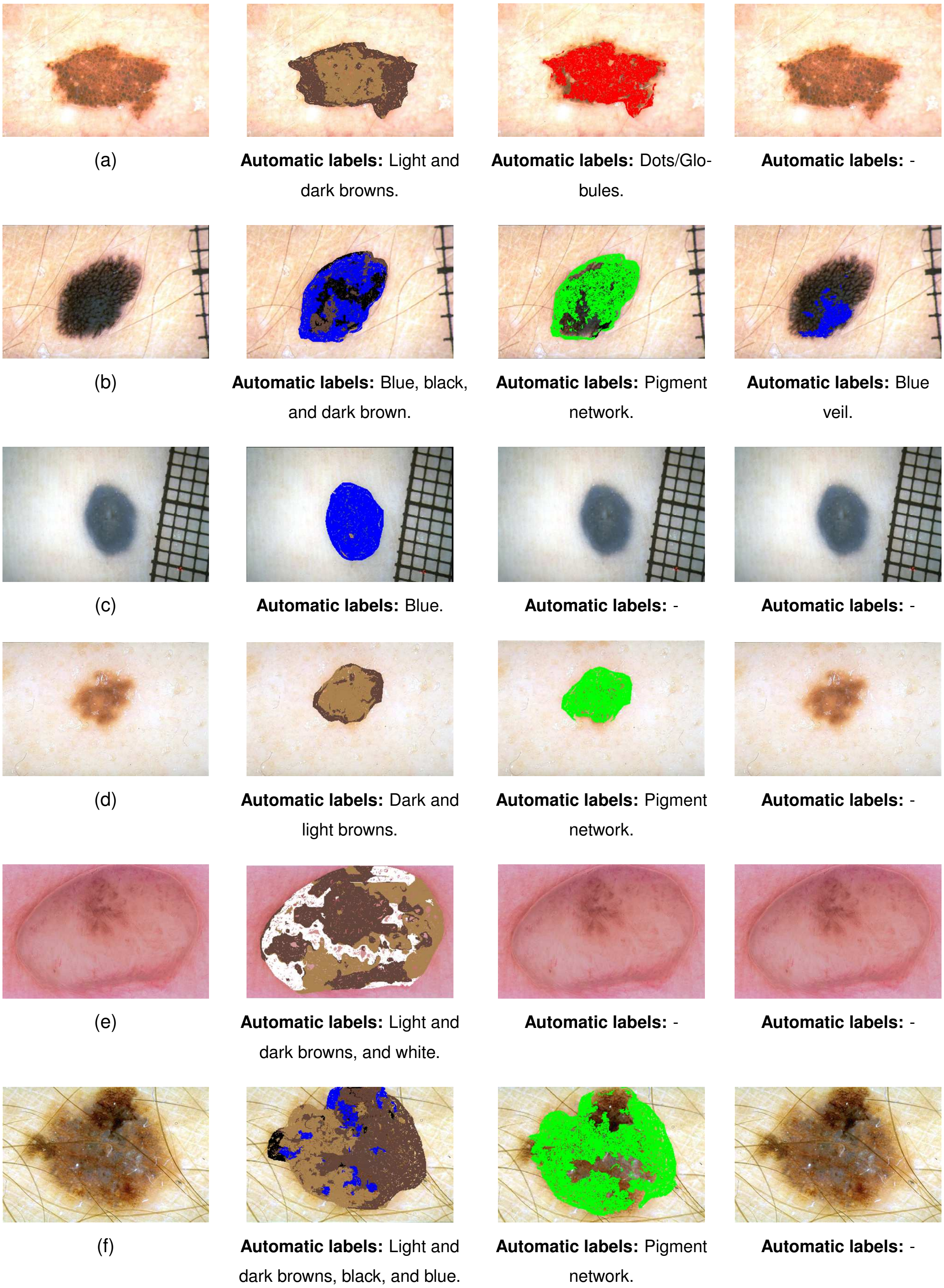
Criteria	Precision	Recall	F1	Best Configuration
Blue-Gray (#226)	87.6%	94.2%	90.8%	Thresholding μ_{HSV}
Dark-Brown (#303)	95.7%	95.7%	95.7%	
Light-Brown (#247)	89.1%	92.7%	90.9%	
Black (#179)	81.5%	88.8%	85.0%	
Red (#31)	79.3%	74.2%	76.7%	
White (#15)	63.6%	93.3%	75.6%	
Pigment Network (#504)	**77.6%**	**88.9%**	**82.9%**	$\mu_g, \sigma_g, \mu_c, \mu_{ca}$ $\mu_M, \sigma_M, \mu_m, \sigma_m$ Random forests
Dots/Globules (#535)	71.8%	83.2%	77.1%	
Blue-Whitish Veil (#178)	75.4%	68.5%	71.9%	$\mu_{HSV}, \mu_c, \mu_{ca}$ Random forests
Regression Areas (#110)	60.8%	51.3%	55.6%	
Pigment Network* (#504)	78.5%	86.1%	82.1%	$\mu_m, \sigma_m,$ $\mu_c, \mu_{ca}, \mu_M, \sigma_M,$ $\mu_{HSV},$ Random forests
Dots/Globules* (#535)	**72.8%**	**83.7%**	**77.9%**	
Blue-Whitish Veil* (#178)	**82.8%**	**68.1%**	**74.7%**	
Regression Areas* (#110)	**63.9%**	**58.8%**	**61.2%**	

those of the other criteria. Interestingly, the best overall results are obtained when we combine all of the structures in a single model (4th row), leading to significant improvements in the detection of blue-whitish veil and regression areas.

Figures 9.6 and 9.7 show some test examples of the automatic annotation and segmentation of medical criteria. Although there is ground-truth segmentation for the criteria (recall that the model was trained using text labels only), the region labeling proposed by the model seems to provide a correct interpretation of the lesion, namely lesions (a-f). Among these results, it is interesting to point out lesion (c), which is a blue-nevus. It is very common for automatic methods to incorrectly associate the blue shade of this type of lesion with blue-whitish veil, as reported in [37, 120]. Nonetheless, the proposed framework is able to successfully distinguish between the blue-nevus and blue-whitish veil. In lesions (g), (i), (h), and (j) the system is able to correctly identify the presence several criteria. However, it also performs the following incorrect annotations:

- **(g) -** pigment network and the color red.
- **(i) -** Color information for this lesion is not available, but a trained eye can easily observe that a red label was incorrectly given to brown regions.
- **(h) -** Color information for this lesion is not available, but a trained eye can easily observe that a light-brown label was incorrectly given to white regions. Moreover, the pigment network label extends to other regions that do not exhibit this structure.
- **(j) -** Similarly to lesion (i), the pigment network and dots/globules annotations are correct but it is observable that they extend to other regions that are not associated with those structures.

The problem of excessive region labeling observed in lesions (h) and (j) is solved when a single model is trained to identify the four structures (see Figure 9.8).



(a)

Automatic labels: Light and dark browns.

Automatic labels: Dots/Globules.

Automatic labels: -

(b)

Automatic labels: Blue, black, and dark brown.

Automatic labels: Pigment network.

Automatic labels: Blue veil.

(c)

Automatic labels: Blue.

Automatic labels: -

Automatic labels: -

(d)

Automatic labels: Dark and light browns.

Automatic labels: Pigment network.

Automatic labels: -

(e)

Automatic labels: Light and dark browns, and white.

Automatic labels: -

Automatic labels: -

(f)

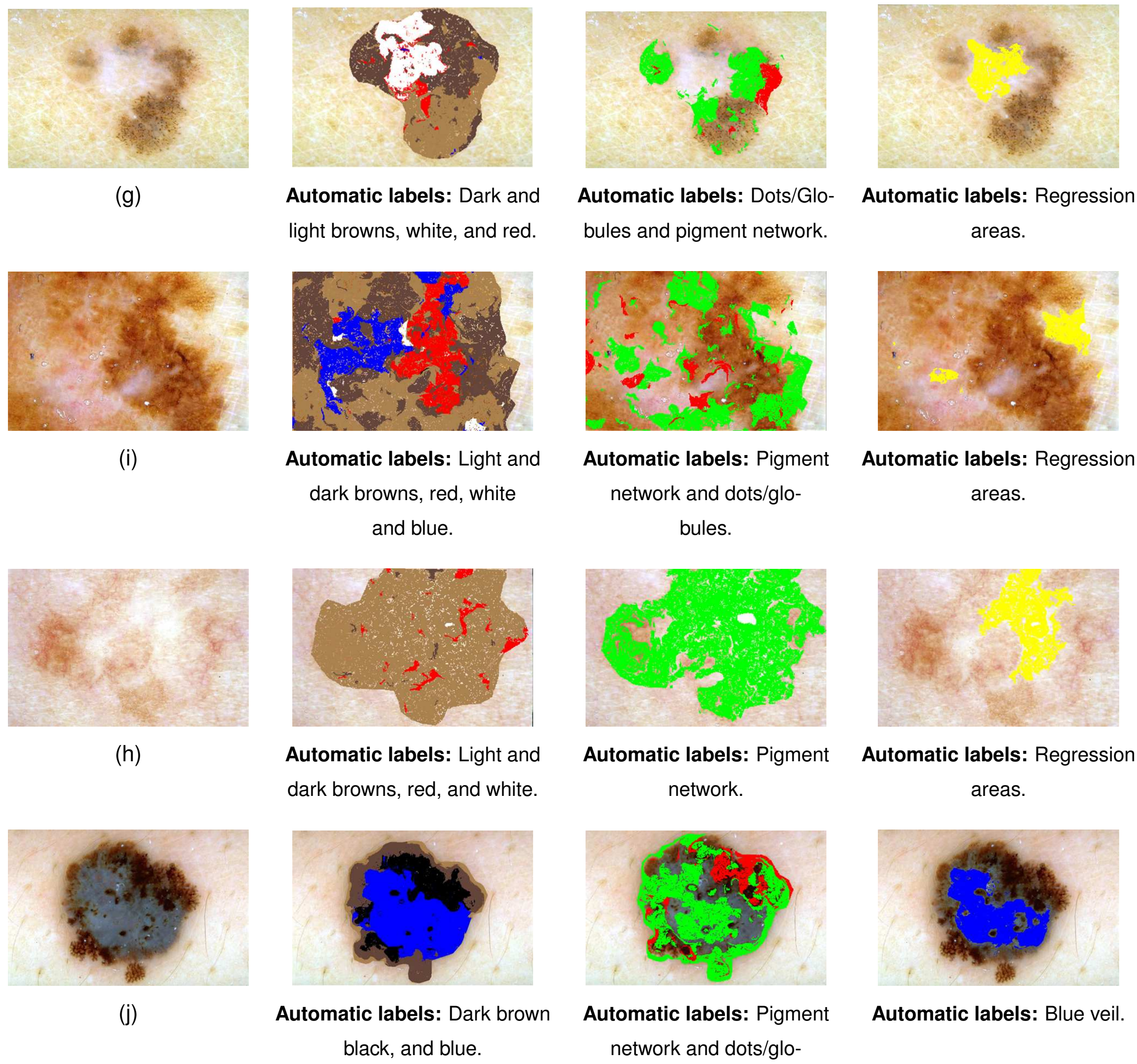
Automatic labels: Light and dark browns, black, and blue.

Automatic labels: Pigment network.

Automatic labels: -

Figure 9.6: Correct detection results: original image (1st column), region and image annotation obtained with the color, texture structures, and color structures models (2nd-4th columns. The color scheme is green for pigment network, red for dots/globules, blue for blue whitish veil, yellow for regression areas.

Table 9.3 compares the performance of the method described in this chapter with the color detection one proposed in Chapter 8. Recall that the latter method was trained using detailed color segmentations, while the method described in this chapter was trained using text labels only. Unfortu-



(g)

Automatic labels: Dark and light browns, white, and red.

Automatic labels: Dots/Globules and pigment network.

Automatic labels: Regression areas.

(i)

Automatic labels: Light and dark browns, red, white and blue.

Automatic labels: Pigment network and dots/globules.

Automatic labels: Regression areas.

(h)

Automatic labels: Light and dark browns, red, and white.

Automatic labels: Pigment network.

Automatic labels: Regression areas.

(j)

Automatic labels: Dark brown black, and blue.

Automatic labels: Pigment network and dots/glo-

Automatic labels: Blue veil.

Figure 9.7: Incorrect detection results: original image (1st column), region and image annotation obtained with the color, texture structures, and color structures models (2nd-4th columns). The color scheme is green for pigment network, red for dots/globules, blue for blue whitish veil, yellow for regression areas.

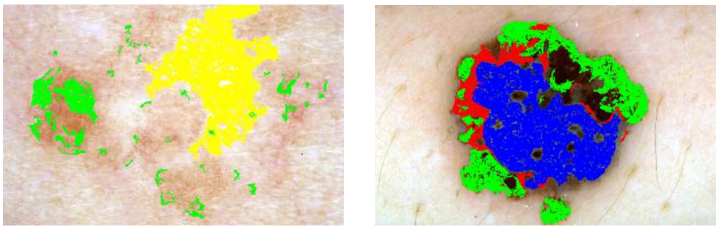


Figure 9.8: Output of the corr-LDA that combines color and texture structures for lesions (h) and (j). The color scheme is green for pigment network, red for dots/globules, blue for blue whitish veil, yellow for regression areas.

nately, as reported in Chapter 8, the number of color segmentations was small and it was not possible to model the red color due to lack of training examples. Nonetheless, it is still possible to compare the performance of the two methods for the remaining colors. Both color detection methods perform well for most of the colors. It is easy to notice that the corr-LDA method outperforms Gaussian mixtures model in the case of the white color. Moreover, it seems to achieve a better recall score for all of the colors. The Gaussian mixtures method achieves a better precision score for three of the colors. Nonetheless, the precision scores obtained using corr-LDA are also high and remarkable, especially if one considers that the Gaussian mixture model was trained with detailed manual segmentations provided by an expert, while the corr-LDA model was trained with text labels only, *i.e.*, without any information about the spatial distribution of the cues.

Table 9.3: Comparison of color detection methods. **bold** highlights the best results.

	corr-LDA		**Gaussian Mixtures 8**	
	Precision	**Recall**	**Precision**	**Recall**
Blue-Gray	**87.6%**	**94.2%**	86.5%	92.2%
Dark-Brown	95.7%	**95.7%**	**98.3%**	76.4%
Light-Brown	89.1%	**92.7%**	**97.0%**	81.0%
Black	81.5%	**88.8%**	**90.9%**	67.0%
Red	**79.3%**	**74.2%**	-	-
White	**63.6%**	**93.3%**	42.1%	85.7%

A comparison with state-of-the-art methods that detect pigment network and blue-whitish veil is shown in Table 9.4). This comparison is possible because all of the methods are trained and tested on images of the same database used in this chapter (EDRA [9]. A comparison with the algorithm proposed in Chapter 3 [16] is also performed. In this case, the method was applied to the same images used in this chapter. The proposed method compares favorably with all of these works. Moreover, the system detects multiple structures, while [11,119,120,159] only focus on one structure, and it is not trained using segmentations of the criteria, as is the case of [119].

Table 9.4: Comparison of structure detection methods. The numbers between parenthesis are the # of images with the criterion and the size of the dataset. ”*” identifies the results obtained with the combined structures model.

Type of Criteria	**Recall**	**Precision**	**F1**	**Method (#Images)**
Pigment Network	83.8%	79.2%	81.4%	[16] (#504/804)
	82.3%	82.3%	82.1%	[159] (#275/436)
	86.0%	79.6%	82.7%	[11](#100/220)
	88.9%	77.6%	82.9%	Proposed (#504/804)
	86.1%*	78.5%*	82.1%*	Proposed* (#504/804)
Blue-whitish veil	70.0%	71.0%	70.5%	[119](#173/223)
	68.1%	63.8%	65.9%	[120](#198/855)
	68.5%	75.4%	71.8%	Proposed(#178/804)
	68.1%*	82.8%*	77.9%*	Proposed*(#178/804)

9.9.3 Lesion Diagnosis

This section evaluates the final decision provided by the system (benign vs melanoma). Three experiments were performed at this stage:

i) Assess the performance of the different medical criteria and their combination.

Table 9.5: Results for melanoma diagnosis. "All*" refers to the combination of the colors with color + texture structures and "Combined" refers to the late fusion of SVM and random forests.

Criteria	SVM	Random Forests	Combined
Colors	$SE = 72.6\%$ $SP = 71.4\%$ $S = 0.279$	$SE = 84.6\%$ $SP = 61.1\%$ $S = 0.248$	$SE = 76.7\%$ $SP = 66.8\%$ $S = 0.273$
Texture Structures	$SE = 87.1\%$ $SP = 60.0\%$ $S = 0.237$	$SE = 75.1\%$ $SP = 75.1\%$ $S = 0.249$	$SE = 81.7\%$ $SP = 66.1\%$ $S = 0.245$
Color Structures	$SE = 83.4\%$ $SP = 70.9\%$ $S = 0.216$	$SE = 82.1\%$ $SP = 72.5\%$ $S = 0.217$	$SE = 84.6\%$ $SP = 73.4\%$ $S = 0.199$
Color Structures & Texture Structures	$SE = 90.9\%$ $SP = 56.0\%$ $S = 0.231$	$SE = 80.9\%$ $SP = 74.8\%$ $S = 0.215$	$SE = 83.4\%$ $SP = 68.1\%$ $S = 0.227$
All	$SE = 84.2\%$ $SP = 72.1\%$ $S = 0.206$	$SE = 81.3\%$ $SP = 74.8\%$ $S = 0.213$	$SE =$ **84.6%** $SP =$ **74.2%** $S =$ **0.196**
All*	$SE = 85.4\%$ $SP = 65.4\%$ $S = 0.226$	$SE = 82.3\%$ $SP = 73.2\%$ $S = 0.213$	$SE =$ **85.8%** $SP =$ **71.1%** $S =$ **0.201**

ii) Compare the diagnostic accuracy of SVM with RBF kernel and random forests.

iii) Compare two late fusion strategies (LR and MR).

All of the experiments were carried on using the outputs of the best corr-LDA models, selected in the previous section. The number of trees for the Random Forests algorithm was set to be $T \in \{1, 3, 5, .., 201\}$, while the parameters of SVM-RBF are set to be $\rho \in \{2^{-12}, 2^{-5}, ..., 2^{12}\}$ and $C \in \{2^{-6}, 2^{-4}, ..., 2^{6}\}$. All of the parameters were selected by nested cross validation. The total number of configurations were: 36360 for random forests and 11700 for SVM. In all of the experiments the features were normalized, as described in Chapter 4.

Table 9.5 shows the best results for all the experiences. The remaining results can be found in Appendix D. As expected, the combination of all the criteria improves the scores. Moreover, due to the characteristics of late fusion, it was also possible to combine the outputs of the two different classifiers (SVM and random forests) [104], leading to the best classification scores. These results were achieved using MR as the fusion strategy. Although MR outperformed LR in all the experiments, the latter also led to relevant classification scores: $SE = 79.2\%$, $SP = 75.6\%$, and $AUC = 83.6\%$. However the results obtained using MR ($SE = 84.6\%$, $SP = 74.2\%$ and $SE = 85.8\%$, $SP = 71.1\%$), are clearly better.

These results compare favorably with the ones obtained in Chapter 7 ($SE = 83.0\%$, $SP = 76.0\%$, and $S = 0.198$), where abstract image features were combined in order to diagnose melanomas. This suggests that it is possible to replace the more traditional pattern recognition features by ones that can be associated with medical information, without losing classification power, provided that there are text labels associated to the training images.

9.10 Conclusions

This chapter discusses the development of a system for skin lesion diagnosis that is inspired by clinical practice of expert dermatologists, and tries to mimic the different steps of medical analysis. First, it divides the lesion into regions that have similar color and texture properties. Then, each of the regions is characterized by a feature vector, and an image annotation algorithm called correspondence-latent Dirichlet allocation is used to label each of them. The labels correspond to the dermoscopic criteria assessed by dermatologists in their diagnostic procedures. Finally, all of the detected criteria are combined in order to obtain a diagnosis.

The proposed computer aided diagnosis system is different from any other found in literature. It is able to deal with the problem of weakly annotated images (text labels), using an image annotation algorithm. This means that it is possible to train the system to detect a dermoscopic criterion without requiring segmentations of that criterion. Moreover, the system is able to detect multiple dermoscopic criteria at the same time. This leads to a medical description of the lesion that can be used to predict a diagnosis. The system is also unique in its ability to combine the information of the different types of medical criteria to provide a decision.

Experimental results for criteria detection are promising and show that the proposed framework can be used to identify the six ABCD rule colors, pigment network, dots/globules, blue-whitish veil, and white regression areas. The system detects not only the presence of the criteria but their spatial localization as well. The detected criteria provide information about different properties of the lesion and are indicative of a malignant lesion. The melanoma detection scores are also promising: sensitivity = 84.6% and specificity = 74.2%. These scores compare favorably with the ones obtained in Chapter 7 using traditional image descriptors. This is an evidence that it is possible to develop systems that use features with a medical meaning, without decreasing the classification performance, and opens a new direction of progress in the analysis of desmoscopic images.

10

Conclusions and Future Work

Contents

10.1 Conclusions

This thesis addressed the problem of melanoma detection using dermoscopy images. Each of the chapters addressed a different question and proposed a methodology to deal with it.

Chapter 1 was an introduction to the problem addressed in this thesis. It motivating the reader to the problem of melanoma diagnosis, followed by the presentation of the different objectives and main contributions of the work developed during the PhD. The chapter concluded with a list of publications. Chapter 2 started by giving an overview of the different types of skin lesions and a description of the medical procedures used to diagnose them. The two types of CAD systems that can be found in literature (pattern recognition based and clinically inspired) were presented in the sequel. The main characteristics of these categories were presented as well as their pros and cons. An assessment of the most relevant CAD system in each category was also performed. Different dermoscopy datasets were also presented.

Chapter 3 described an algorithm for the detection of pigment network, which is considered by many dermatologists (including the one who collaborated in a significant part of this work) as one of the most relevant dermoscopic cues. This structure is simultaneously used to distinguish between melanocytic and non-melanocytic lesions, as well as to diagnose melanomas. The proposed algorithm detected pigment network using two of its properties: its shape (lines with multiple orientations) and spatial arrangement (a network of connected segments). A bank of directional filters was used to enhance the lines of the network, making use of its linear shape, followed by a connected component analysis and area thresholding, in order to select large regions. Finally, AdaBoost was used to classify the lesions according to the presence or absence of network, achieving a sensitivity of 91.1% and a specificity of 82.1%, on a dataset of 200 images.

The following chapters focused on the detection of melanomas. Chapter 4 investigated the relevance of four types of features inspired by the ABCD rule of dermoscopy: color, texture, shape, and symmetry. Moreover, different descriptors associated to each type of feature were also compared. A study of this kind was missing in literature, where it is common practice to compute multiple descriptors to represent each of the aforementioned features, without providing any insightful information about their discriminative power and even how to compute them. The obtained results showed that color features outperformed the remaining ones, while shape features led to the poorest performances.

The descriptors used in Chapter 4 are called global features. This type of features is suitable to characterize some of the elements of the ABCD rule. However, it is debatable if global features are appropriate to characterize localized dermoscopic criteria. In order to answer to this question, Chapter 5 investigated the role of local features, using the BoF model to separately characterize small patches extracted from the lesions. The performance of these features was compared against that of the global ones. It was shown that local features performed well and that it was better to characterize the texture of the lesions using these features. This thesis was one of the first works to propose the use of local features in the context of dermoscopy image analysis.

The experiments performed in Chapters 4 and 5 where carried on using a dataset acquired with a single device and specific illumination conditions. Changes in one of these aspects may degrade the performance of the system, since there will some color variability between the images used to train the system and the new ones. This issue was addressed in Chapter 6, where a strategy to deal with images acquired at multiple hospitals was studied. In this chapter four simple color normalization algorithms were applied to a multi-source dataset and it was determined that all of them increased the accuracy of the CAD system, with shades of gray being the best method. It is important to point out that all of the investigated methods are very simple and do not require any training nor knowledge about the acquisition system, since they normalize the colors of the images using simple image statistics.

After separately studying each of the four types of features (color, texture, shape, and symmetry), assessing different descriptors for each of them, and comparing global and local feature extraction processes, it was necessary to combine all of the information into a single CAD system. This problem was investigated in Chapter 7, where two methodologies for feature fusion were compared: early and late fusion. The experiments showed that late fusion outperformed early fusion.

Chapters 4 to 7 comprise the stage of this thesis that was related with traditional pattern recognition approaches for melanoma diagnosis. The experiments performed in these chapters provided insightful information about: i) the role of each feature; ii) the best descriptors; iii) the importance of local features; iv) the optimal strategy to combine different features into a single CAD system; and v) simple strategies to increase the robustness of the CAD system to different acquisition conditions. All of the experiments are fully reported in several papers, thus it is possible to reproduce any of them. Moreover, the work developed in Chapters 4, 5, and 7 was conducted using the publicly available PH^2 dataset, meaning that the obtained results can be used as a baseline for comparison by other works.

The last stage of this thesis was the development of a clinically inspired CAD system that fulfilled the needs of the medical community. The goal of the system was to provide comprehensive information to justify the automatic diagnosis of the lesions. This was accomplished by mimicking the medical way of diagnosing skin lesions, *i.e.*, by extracting features with clinical meaning, followed by lesion diagnosis based on those features. The first step towards this goal was the development of a strategy to detect clinically relevant colors in dermoscopy images using Gaussian mixture models (Chapter 8). This algorithm achieved promising results, but had a limitation: it requires training examples with detailed annotations (text labels and segmentations of criteria performed by experts), which are very difficult to obtain for all the medical criteria. Unfortunately, such requirement makes it impossible to generalize the method to other dermoscopic criteria. The lack of training examples with detailed annotations is a major limitation in the development of any clinically inspired system, since dermatologists consider the annotation task time consuming and do not perform it. Chapter 9 described a strategy to deal with weakly annotated dermoscopy images, *i.e.*, the annotations consist only of text labels, with no information about the location of the dermoscopic criteria. A probabilistic image annotation algorithm was applied to this problem, making it possible for the system to learn to detect multiple dermoscopic criteria, without relying on segmentations. The proposed system was able to detect several dermoscopic criteria that are relevant for melanoma diagnosis. These criteria were then used to train a classification algorithm to diagnose dermoscopy images, achieving very good scores. Moreover, the obtained statistics were comparable with the ones achieved using the pattern recognition methods with regular image processing features, investigated this thesis. This led to the conclusion that it is possible to develop systems that are able to provide a diagnosis using features that have a medical meaning, making it possible for the dermatologists to understand and validate the diagnosis. The proposed system provides a significant advance in dermoscopy image analysis, since it is able to: i) deal with poorly annotated training data; ii) detect multiple structures at the same time; and iii) provide a diagnosis using medical information.

10.2 Future Work

This document reports a journey that can be divided into two mains stages: i) a thorough study of pattern recognition approaches for melanoma diagnosis; and ii) a road towards the development of a clinically inspired CAD system. Both stages resulted in relevant outputs for the research community. Based on the outputs and on the knowledge acquired during the development of this thesis, it is possible to give some guidelines for future work on this fields.

i) **Pattern recognition methods:** This thesis has applied state-of-the-art classifiers, but it is undeniable that deep learning has been outperforming other classification algorithm in several classification tasks (*e.g.* [45, 109, 201]). One of the main reasons for the ever increasing interest in neural networks is the amount of data that is nowadays available to train the models. However, dermoscopy datasets tend to be small: the largest datasets do not contain more than 2500 images (*e.g.*, EDRA [9] or the very recent ISIC dataset [82]), which might make them unfit to train a deep neural network that requires the optimization of a very high number of hyperparameters. This poses the question of how to deal with small datasets if one wants to use deep learning. Is it possible to apply a pre-trained network to dermoscopy images and obtain good classification scores? These networks are trained using completely different images, thus they may not be appropriate for dermoscopy images. Therefore, it is also important to access what is the minimum amount of data that is necessary to completely train a deep neural network.

   Another interesting research topic would be to test other strategies to extract local features, namely sparse coding (SC), which has been shown to outperform BoF in some applications [200, 204]. The main strength of (SC) is that it relaxes the constraint imposed by BoF that each patch is only associated with one of the elements of the dictionary, assuming instead that the patch is a combination of a small number of elements. This makes SC more flexible and, theoretically, more efficient in describing the images [122]. Therefore, it is important to determine if applying SC to compute the local features significantly improves the scores of a CAD system. Moreover, the SC framework makes it possible to estimate different types of dictionaries, namely estimate the dictionaries using both benign and malignant lesions or estimate two dictionaries, one for each class, called discriminative dictionaries. This is an interesting line of research, since discriminative dictionaries are more suitable to deal with unbalanced classes, as is the case of melanomas when compared with benign lesions.

ii) **Clinically inspired methods:** The system described in Chapter 9 opens a new direction in dermoscopy image analysis, and it would be very interesting for other works to follow this path. A question that easily comes to mind is the following: is it possible to use other methods to detect the dermoscopic criteria and achieve similar or even better results? Chapter 9 presented several strategies to perform image annotation, some more suitable than others (*e.g.*, the MIML framework). The choice of corr-LDA was justified, but other annotations methods should be tested and compared with corr-LDA. Once again it is possible to consider the use of deep neural networks, which have been used not only to classify the data, but also to provide automatic captions and localized objects in images (*e.g.*, [114,137]). Another interesting question is if there is a limit for the number of criteria that the proposed method is able to detect. The proposed model was able to detect six colors and four structures, but there are still other dermoscopic criteria to be detected. Some of these criteria are specific for melanocytic lesions, while others are associated with non-melanocytic lesions. Therefore, adding additional criteria to the model may be a challenge in terms of training data. Moreover, adding new criteria may require the use of new descriptors to characterize the regions or even new image splitting methods, in order to achieve better regions.

    Other relevant aspect of clinically inspired methods is the type of training set. The one used in this thesis belongs to the category of weakly annotated training sets, which means that only text labels are available. An interesting work would be to understand the influence of using a heterogeneous dataset, *i.e.*, a dataset that also comprised segmentations of the criteria for some of the images. Theoretically, adding information about the location of the criteria would make it easier for the model to learn the appearance of the criteria. However, if the segmentations are few, the information provided by them might not be sufficient. Another possible type of dataset would be the one that comprised three types of images: with detailed annotations, with weak annotations, and no annotations at all. Dealing with training images without annotations would turn the problem described in this thesis into one of semi-supervised learning. Finally, still on the dataset topic, it would be important to investigate the how to deal with cases where the annotations are wrong. Incorrect annotations hamper the learning process, which means that the model will not be properly trained. It will be necessary to find a strategy to deal with wrong annotations, such as a assigning a degree of confidence to each expert.

All of the experiments from Chapters 4 to 9 were performed using datasets that only comprised melanocytic lesions. Dermatologists acquire images of both melanocytic and non-melanocytic lesions, but the latter have been ignored by most of the research community. Therefore, it is still important to develop strategies to deal with datasets that contain both types of images. This is something that is common to both the pattern recognition and the clinically inspired CAD systems.

Something that was attempted during the development of this thesis but that was not possible to carry on due several constraints was the extension of the developed methods to the real world tasks.

In other words, the developed CAD systems should be tested in real time during the routine clinical exams. This kind of experiment is also missing in the literature and involves the participation of one or more dermatology services.

Bibliography

[1] "Dermnet," http://www.dermnet.com/.

[2] "Dermoscopy atlas," http://www.dermoscopyatlas.com.

[3] Q. Abbas, M. Celebi, and I. Fondón, "Computer-aided pattern classification system for dermoscopy images," Skin Research and Technology, vol. 18, no. 3, pp. 278–289, 2012.

[4] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, no. 11, pp. 2274–2282, 2012.

[5] N. Alfed, F. Khelifi, A. Bouridane, and H. Seker, "Pigment network-based skin cancer detection," in 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, 2015, pp. 7214–7217.

[6] M. Anantha, R. Moss, and W. Stoecker, "Detection of pigment network in dermatoscopy images using texture analysis," Computarized Medical Imaging and Graphics, vol. 28, no. 5, pp. 225–234, 2004.

[7] G. Argenziano, G. Fabbrocini, P. Carli, V. De Giorgi, E. Sammarco, and E. Delfino, "Epiluminescence microscopy for the diagnosis of doubtful melanocytic skin lesions. comparison of the ABCD rule of dermatoscopy and a new 7-point checklist based on pattern analysis," Archives of Dermatology, vol. 134, pp. 1563–1570, 1998.

[8] G. Argenziano, H. P. Soyer, S. Chimenti, R. Talamini, R. Corona, F. Sera, M. Binder, L. Cerroni, G. De Rosa, and G. Ferrara, "Dermoscopy of pigmented skin lesions: results of a consensus meeting via the internet," Journal of the American Academy of Dermatology, vol. 48, no. 5, pp. 679–693, 2003.

[9] G. Argenziano, H. P. Soyer, V. De Giorgi, D. Piccolo, P. Carli, M. Delfino, A. Ferrari, V. Hofmann-Wellenhog, D. Massi, G. Mazzocchetti, M. Scalvenzi, and I. H. Wolf, Interactive Atlas of Dermoscopy. EDRA Medical Publishing & New Media, 2000.

[10] S. Arivazhagan, L. Ganesan, and S. Priyal, "Texture classification using gabor wavelets based rotation invariant features," Pattern Recognition Letters, vol. 27, no. 16, pp. 1976–1982, 2006.

[11] J. Arroyo and B. G. Zapirain, “Automated detection of melanoma in dermoscopic images,” in Computer Vision Techniques for the Diagnosis of Skin Cancer, J. Scharcanski and M. E. Celebi, Eds. Springer, 2014, pp. 139–192.

[12] C. Barata, M. E. Celebi, and J. S. Marques, “Color detection in dermoscopy images based on scarce annotations,” in Iberian Conference on Pattern Recognition and Image Analysis. Springer, 2015, pp. 309–316.

[13] ——, “Improving dermoscopy image classification using color constancy,” IEEE Journal of Biomedical and Health Informatics, vol. 19, no. 3, pp. 1146–1152, May 2015.

[14] C. Barata, M. E. Celebi, J. S. Marques, and J. Rozeira, “Clinically inspired analysis of dermoscopy images using a generative model,” Computer Vision and Image Understanding, vol. 151, pp. 124–137, 2016.

[15] C. Barata, J. S. Marques, and M. E. Celebi, “Improving dermoscopy image analysis using color constancy,” in IEEE International Conference on Image Processing (ICIP). IEEE, 2014, pp. 3527–3531.

[16] C. Barata, J. Marques, and J. Rozeira, “A system for the detection of pigment network in dermoscopy images using directional filters,” IEEE Transactions on Biomedical Engineering, vol. 59, no. 10, pp. 2744–2754, 2012.

[17] ——, “Evaluation of color based keypoints and features for the classification of melanomas using the bag-of-features model,” in Advances in Visual Computing. Springer, 2013, pp. 40–49.

[18] ——, “The role of keypoint sampling on the classification of melanomas in dermoscopy images using bag-of-features,” in Pattern Recognition and Image Analysis. Springer, 2013, pp. 715–723.

[19] C. Barata, M. Ruela, M. Francisco, T. Mendonça, and J. Marques, “Two systems for the detection of melanomas in dermoscopy images using texture and color features,” IEEE Systems Journal, vol. 8, no. 3, pp. 965–979, 2014.

[20] C. Barata, M. Ruela, T. Mendonça, and J. Marques, “A bag-of-features approach for the classification of melanomas in dermoscopy images: The role of color and texture descriptors,” in Computer Vision Techniques for the Diagnosis of Skin Cancer. Springer, 2014, pp. 49–69.

[21] K. Barnard, P. Duygulu, D. Forsyth, N. De Freitas, D. Blei, and M. Jordan, “Matching words and pictures,” The Journal of Machine Learning Research, vol. 3, pp. 1107–1135, 2003.

[22] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester, “Image inpainting,” in Proceedings of the 27th annual conference on Computer graphics and interactive techniques. ACM Press/Addison-Wesley Publishing Co., 2000, pp. 417–424.

[23] G. Betta, G. Di Leo, G. Fabbrocini, A. Paolillo, and P. Sommella, “Dermoscopic image-analysis system: estimation of atypical pigment network and atypical vascular pattern,” in 2006 IEEE International Workshop on Medical Measurement and Applications (MeMea). IEEE, 2006, pp. 63–67.

[24] M. Binder, M. Puespoeck-Schwarz, A. Steiner, H. Kittler, M. Muellner, K. Wolff, and H. Pehamberger, “Epiluminescence microscopy of small pigmented skin lesions: short-term formal training improves the diagnostic performance of dermatologists,” Journal of the American Academy of Dermatology, vol. 36, no. 2, pp. 197–202, 1997.

[25] D. Blei and M. Jordan, “Modeling annotated data,” in 26th annual international ACM SIGIR conference on Research and development in informaion retrieval. ACM, 2003, pp. 127–134.

[26] D. Blei, A. Ng, and M. Jordan, “Latent dirichlet allocation,” the Journal of machine Learning research, vol. 3, pp. 993–1022, 2003.

[27] M. R. Boutell, J. Luo, X. Shen, and C. M. Brown, “Learning multi-label scene classification,” Pattern recognition, vol. 37, no. 9, pp. 1757–1771, 2004.

[28] M. Bratkova, S. Boulos, and P. Shirley, “orgb: a practical opponent color space for computer graphics.” IEEE Computer Graphics and Applications, vol. 29, no. 1, pp. 42–55, 2009.

[29] L. Breiman, “Random forests,” Machine learning, vol. 45, no. 1, pp. 5–32, 2001.

[30] G. Buchsbaum, “A spatial processor model for object colour perception,” Journal of the Franklin institute, vol. 210, pp. 1–26, 1980.

[31] M. Burroni, R. Corona, G. Dell’Eva, F. Sera, R. Bono, P. Puddu, R. Perotti, F. Nobile, L. Andreassi, and P. Rubegni, “Melanoma computer-aided diagnosis reliability and feasibility study,” Clinical cancer research, vol. 10, no. 6, pp. 1881–1886, 2004.

[32] S. Calderara, A. Prati, and R. Cucchiara, “Mixtures of von mises distributions for people trajectory shape analysis,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 21, no. 4, pp. 457–471, 2011.

[33] G. Capdehourat, A. Corez, A. Bazzano, R. Alonso, and P. Musé, “Toward a combined tool to assist dermatologists in melanoma detection from dermoscopic images of pigmented skin lesions,” Pattern Recognition Letters, vol. 32, no. 16, pp. 2187–2196, 2011.

[34] G. Carneiro, A. Chan, P. Moreno, and N. Vasconcelos, “Supervised learning of semantic classes for image annotation and retrieval,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 29, no. 3, pp. 394–410, 2007.

[35] C. Carson, S. Belongie, H. Greenspan, and J. Malik, “Blobworld: Image segmentation using expectation-maximization and its application to image querying,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 8, pp. 1026–1038, 2002.

[36] M. E. Celebi, H. Iyatomi, G. Schaefer, and W. V. Stoecker, "Lesion border detection in dermoscopy images," Computerized medical imaging and graphics, vol. 33, no. 2, pp. 148–153, 2009.

[37] M. E. Celebi, H. Iyatomi, W. V. Stoecker, R. H. Moss, H. S. Rabinovitz, G. Argenziano, and H. P. Soyer, "Automatic detection of blue-white veil and related structures in dermoscopy images," Computerized Medical Imaging and Graphics, vol. 32, pp. 670–677, 2008.

[38] M. E. Celebi, H. Kingravi, B. Uddin, H. Iyatomi, Y. A. Aslandogan, W. Stoecker, and R. Moss, "A methodological approach to the classification of dermoscopy images," Computerized Medical Imaging and Graphics, vol. 31, no. 6, pp. 362–373, 2007.

[39] M. E. Celebi, Q. Wen, S. Hwang, and G. Schaefer, "Color quantization of dermoscopy images using the k-means clustering algorithm," in Color Medical Image Analysis. Springer, 2013, pp. 87–107.

[40] M. E. Celebi, Q. Wen, H. Iyatomi, K. Shimizu, H. Zhou, and G. Schaefer, "A state-of-the-art survey on lesion border detection in dermoscopy images," M. E. Celebi, T. Mendonça, and J. S. Marques, Eds. CRC Press, 2015, pp. 97–129.

[41] M. E. Celebi and A. Zornberg, "Automated quantification of clinically significant colors in dermoscopy images and its application to skin lesion classification," IEEE Systems Journal, vol. 8, no. 3, pp. 980–984, 2014.

[42] G. Celeux, S. Chrétien, F. Forbes, and A. Mkhadri, "A component-wise em algorithm for mixtures," Journal of Computational and Graphical Statistics, vol. 10, 2001.

[43] W. Cheng and E. Hüllermeier, "Combining instance-based learning and logistic regression for multilabel classification," Machine Learning, vol. 76, no. 2-3, pp. 211–225, 2009.

[44] A. Clare and R. D. King, "Knowledge discovery in multi-label phenotype data," in Principles of data mining and knowledge discovery. Springer, 2001, pp. 42–53.

[45] N. Codella, J. Cai, M. Abedini, R. Garnavi, A. Halpern, and J. R. Smith, "Deep learning, sparse coding, and svm for melanoma recognition in dermoscopy images," in International Workshop on Machine Learning in Medical Imaging. Springer, 2015, pp. 118–126.

[46] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 5, pp. 603–619, 2002.

[47] C. Cortes and V. Vapnik, "Support-vector networks," Machine learning, vol. 20, no. 3, pp. 273–297, 1995.

[48] I. Daftari, E. Aghaian, J. M. OBrien, W. Dillon, and T. L. Phillips, "3d mri-based tumor delineation of ocular melanoma and its comparison with conventional techniques," Medical physics, vol. 32, no. 11, pp. 3355–3362, 2005.

[49] A. Dalal, R. Moss, R. Stanley, W. Stoecker, K. Gupta, D. Calcara, J. Xu, B. Shrestha, R. Drugge, and J. Malters, "Concentric decile segmentation of white and hypopigmented areas in dermoscopy images of skin lesions allows discrimination of malignant melanoma," Computerized Medical Imaging and Graphics, vol. 35, no. 2, pp. 148–154, 2011.

[50] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1. IEEE, 2005, pp. 886–893.

[51] J. G. Daugman, "Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters," Journal of the Optical Society of America, vol. 2, no. 7, pp. 1160–1169, 1985.

[52] G. Day and R. Barbour, "Automated skin lesion screening–a new approach," Melanoma research, vol. 11, no. 1, pp. 31–35, 2001.

[53] G. R. Day and R. H. Barbour, "Automated melanoma diagnosis: where are we at?" Skin Research and Technology, vol. 6, no. 1, pp. 1–5, 2000.

[54] O. Debeir, C. Decaestecker, J. Pasteels, I. Salmon, R. Kiss, and P. Van Ham, "Computer-assisted analysis of epiluminescence microscopy images of pigmented skin lesions," Cytometry, vol. 37, no. 4, pp. 255–266, 1999.

[55] Y. Deng and B. Manjunath, "Unsupervised segmentation of color-texture regions in images and video," IEEE transactions on pattern analysis and machine intelligence, vol. 23, no. 8, pp. 800–810, 2001.

[56] G. Di Leo, G. Fabbrocini, A. Paolillo, O. Rescigno, and P. Sommella, "Towards an automatic diagnosis system for skin lesions: estimation of blue-whitish veil and regression structures," in Systems, Signals and Devices, 2009. SSD'09. 6th International Multi-Conference on. IEEE, 2009, pp. 1–6.

[57] G. Di Leo, A. Paolillo, P. Sommella, and G. Fabbrocini, "Automatic diagnosis of melanoma: a software system based on the 7-point check-list," in 2010 43rd Hawaii International Conference on System Sciences (HICSS). IEEE, 2010, pp. 1–10.

[58] M. Douze, D. Oneata, M. Paulin, C. Leray, N. Chesneau, D. Potapov, J. Verbeek, K. Alahari, Z. Harchaoui, and L. Lamel, "The inria-lim-vocr and axes submissions to trecvid 2014 multimedia event detection," 2014.

[59] S. Dreiseitl and M. Binder, "Do physicians value decision support? a look at the effect of decision support systems on physician opinion," Artificial intelligence in medicine, vol. 33, no. 1, pp. 25–30, 2005.

[60] R. Duda, P. Hart, and D. Stork, Pattern classification. John Wiley & Sons, 2012.

[61] P. Duygulu, K. Barnard, J. de Freitas, and D. Forsyth, “Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary,” in European Conference on Computer Vision. Springer, 2002, pp. 97–112.

[62] A. Elisseeff and J. Weston, “A kernel method for multi-labelled classification,” in Advances in neural information processing systems, 2001, pp. 681–687.

[63] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, “Object detection with discriminatively trained part-based models,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, no. 9, pp. 1627–1645, 2010.

[64] P. F. Felzenszwalb and D. P. Huttenlocher, “Efficient graph-based image segmentation,” International Journal of Computer Vision, vol. 59, no. 2, pp. 167–181, 2004.

[65] S. L. Feng, R. Manmatha, and V. Lavrenko, “Multiple bernoulli relevance models for image and video annotation,” in Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2. IEEE, 2004, pp. II–1002.

[66] J. Ferlay, E. Steliarova-Foucher, J. Lortet-Tieulent, S. Rosso, J. Coebergh, H. Comber, D. Forman, and F. Bray, “Cancer incidence and mortality patterns in Europe: estimates for 40 countries in 2012,” European journal of cancer, vol. 49, no. 6, pp. 1374–1403, 2013.

[67] M. A. T. Figueiredo and A. K. Jain, “Unsupervised learning of finite mixture models,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 3, pp. 381–396, 2002.

[68] G. Finlayson, S. Hordley, and P. Hubel, “Color by correlation: A simple, unifying framework for color constancy,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, no. 11, pp. 1209–1221, 2001.

[69] G. Finlayson and E. Trezzi, “Shades of gray and colour constancy,” in In Proceedings of IS&T/SID 12th Color Imaging Conference, 2004, pp. 37–41.

[70] M. Fleming, C. Steger, J. Zhang, J. Gao, A. Cognetta, I. Pollak, and C. Dyer, “Techniques for a structural analysis of dermatoscopic imagery,” Computarized Medical Imaging and Graphics, no. 5, pp. 375–389, 1998.

[71] D. Forsyth, “A novel algorithm for color constancy,” International Journal of Computer Vision, vol. 5, no. 1, pp. 5–36, 1990.

[72] Y. Freund and R. Schapire, “A decision-theoretic generalization of on-line learning and an application to boosting,” in Computational learning theory. Springer, 1995, pp. 23–37.

[73] J. Frühauf, B. Leinweber, R. Fink-Puches, V. Ahlgrimm-Siess, E. Richtig, I. Wolf, A. Niederkorn, F. Quehenberger, and R. Hofmann-Wellenhof, “Patient acceptance and diagnostic utility of automated digital image analysis of pigmented skin lesions,” Journal of the European Academy of Dermatology and Venereology, vol. 26, no. 3, pp. 368–372, 2012.

[74] J. Fürnkranz, E. Hüllermeier, E. L. Mencía, and K. Brinker, “Multilabel classification via calibrated label ranking,” Machine learning, vol. 73, no. 2, pp. 133–153, 2008.

[75] R. Garnavi, M. Aldeen, and J. Bailey, “Classification of melanoma lesions using wavelet-based texture analysis,” in International Conference on Digital Image Computing: Techniques and Applications (DICTA),. IEEE, 2010, pp. 75–81.

[76] N. Ghamrawi and A. McCallum, “Collective multi-label classification,” in Proceedings of the 14th ACM international conference on Information and knowledge management. ACM, 2005, pp. 195–200.

[77] A. Gijsenij, T. Gevers, and J. van de Weijer, “Computational color constancy: Survey and experiments,” IEEE Transactions on Image Processing, vol. 20, no. 9, pp. 2475–2489, 2011.

[78] S. Godbole and S. Sarawagi, “Discriminative methods for multi-labeled classification,” in Advances in Knowledge Discovery and Data Mining. Springer, 2004, pp. 22–30.

[79] C. Grana, R. Cucchiara, G. Pellacani, and S. Seidenari, “Line detection and texture characterization of network patterns.” in In ICPR’06: Proceedings of the 18th International Conference on pattern Recognition, 2006.

[80] C. Grana, G. Pellacani, and S. Seidanari, “Pratical color calibration for dermoscopy applied to a digital epiluminescence microscope,” Skin Research and Technology, vol. 11, pp. 242–247, 2005.

[81] C. Grigorescu, N. Petkov, and M. Westenberg, “Contour detection based on nonclassical receptive field inhibition,” IEEE Transactions on Image Processing, vol. 12, no. 7, pp. 729–739, 2003.

[82] D. Gutman, N. Codella, M. E. Celebi, B. Helba, M. Marchetti, N. Mishra, and A. Halpern, “Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (isbi) 2016, hosted by the international skin imaging collaboration (isic),” arXiv preprint arXiv:1605.01397, 2016.

[83] Y. V. Haeghen, J. M. A. D. Naeyaert, and I. Lemahieu, “An imaging system with calibrated color image acquisition for use in dermatology,” IEEE Transactions on Medical Imaging, vol. 19, no. 7, pp. 722–730, 2000.

[84] R. Haralick, K. Shanmugam, and I. Dinstein, “Textural features for image classification,” IEEE Transactions on Systems, Man and Cybernetics, no. 6, pp. 610–621, 1973.

[85] K. Hoffmann, T. Gambichler, A. Rick, M. Kreutz, M. Anschuetz, T. Grünendick, A. Orlikov, S. Gehlen, R. Perotti, and L. Andreassi, “Diagnostic and neural analysis of skin cancer (danaos). a multicentre study for collection and computer-aided analysis of data from pigmented skin lesions using digital dermoscopy,” British Journal of Dermatology, vol. 149, no. 4, pp. 801–809, 2003.

[86] R. Hong, M. Wang, Y. Gao, D. Tao, X. Li, and X. Wu, “Image annotation by multiple-instance learning with discriminative feature mapping and selection,” IEEE Transactions on Cybernetics, vol. 44, no. 5, pp. 669–680, 2014.

[87] M. Hu, “Visual pattern recognition by moment invariants,” IRE Transactions on Information Theory, vol. 8, no. 2, pp. 179–187, 1962.

[88] H. Iyatomi, M. E. Celebi, G. Schaefer, and M. Tanaka, “Automated color calibration method for dermoscopy images,” Computerized Medical Imaging and Graphics, vol. 35, pp. 89–98, 2011.

[89] H. Iyatomi, K. Norton, M. Celebi, G. Schaefer, M. Tanaka, and K. Ogawa, “Classification of melanocytic skin lesions from non-melanocytic lesions,” in 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, 2010, pp. 5407–5410.

[90] H. Iyatomi, H. Oka, M. E. Celebi, M. Hashimoto, M. Hagiwara, M. Tanaka, and K. Ogawa, “An improved internet-based melanoma screening system with dermatologist-like tumor area extraction algorithm,” Computerized Medical Imaging and Graphics, vol. 32, no. 7, pp. 566–579, 2008.

[91] H. Iyatomi, H. Oka, M. E. Celebi, K. Ogawa, G. Argenziano, H. P. Soyer, H. Koga, T. Saida, K. Ohara, and M. Tanaka, “Computer-based classification of dermoscopy images of melanocytic lesions on acral volar skin,” Journal of Investigative Dermatology, vol. 128, no. 8, pp. 2049–2054, 2008.

[92] A. Jain, Fundamentals of digital image processing. Prentice-Hall, Inc., 1989.

[93] J. Jeon, V. Lavrenko, and R. Manmatha, “Automatic image annotation and retrieval using cross-media relevance models,” in Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval. ACM, 2003, pp. 119–126.

[94] S. Ji, L. Tang, S. Yu, and J. Ye, “Extracting shared subspace for multi-label classification,” in Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2008, pp. 381–389.

[95] Y. Jiang, C. Ngo, and J. Yang, “Towards optimal bag-of-features for object categorization and semantic video retrieval,” in 6th ACM international conference on Image and video retrieval. ACM, 2007, pp. 494–501.

[96] I. Jolliffe, Principal component analysis. Springer, 2002.

[97] J. P. Jones and L. A. Palmer, “An evaluation of the two-dimensional gabor filter model of simple receptive fields in cat striate cortex,” Journal of neurophysiology, vol. 58, no. 6, pp. 1233–1258, 1987.

[98] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, “An introduction to variational methods for graphical models,” Machine learning, vol. 37, no. 2, pp. 183–233, 1999.

[99] P. Kakumanu, S. Makrogiannis, and N. Bourbakis, “A survey of skin-color modeling and detection methods,” Pattern recognition, vol. 40, no. 3, pp. 1106–1122, 2007.

[100] J. Kittler, M. Hatef, R. W. Duin, and J. Matas, “On combining classifiers,” IEEE transactions on pattern analysis and machine intelligence, vol. 20, no. 3, pp. 226–239, 1998.

[101] G. Klinker, S. Shafer, and T. Kanade, “A physical approach to color image understanding.” International Journal of Computer Vision, vol. 4, no. 1, pp. 7–38, 1990.

[102] A. W. Kopf, M. Elbaum, and N. Provost, “The use of dermoscopy and digital imaging in the diagnosis of cutaneous malignant melanoma,” Skin Research and Technology, vol. 3, no. 1, pp. 1–7, 1997.

[103] K. Korotkov and R. Garcia, “Computerized analysis of pigmented skin lesions: a review,” Artificial intelligence in medicine, vol. 56, no. 2, pp. 69–90, 2012.

[104] L. Kuncheva, Combining Pattern Classifiers: Methods and Algorithms. Wiley, 2004.

[105] E. Land, “The retinex theory of color vision,” Scientific America, vol. 237, pp. 108–128, 1977.

[106] V. Lavrenko, R. Manmatha, and J. Jeon, “A model for learning the semantics of pictures,” in Advances in neural information processing systems, 2003, p. None.

[107] K. I. Laws, “Rapid texture identification,” in 24th Annual Technical Symposium. International Society for Optics and Photonics, 1980, pp. 376–381.

[108] S. Lazebnik, C. Schmid, and J. Ponce, “Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories,” in 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2. IEEE, 2006, pp. 2169–2178.

[109] Y. Lequan, H. Chen, Q. Dou, J. Qin, and P. A. Heng, “Automated melanoma recognition in dermoscopy images via very deep residual networks,” IEEE Transactions on Medical Imaging, 2016.

[110] F. Li and C. Sminchisescu, “Convex multiple-instance learning by estimating likelihood ratio,” in Advances in Neural Information Processing Systems, 2010, pp. 1360–1368.

[111] J. Li and J. Wang, “Real-time computerized annotation of pictures,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 30, no. 6, pp. 985–1002, 2008.

[112] M. Lingala, R. Stanley, R. Rader, J. Hagerty, H. Rabinovitz, M. Oliviero, I. Choudhry, and W. Stoecker, “Fuzzy logic color detection: Blue areas in melanoma dermoscopy images,” Computerized Medical Imaging and Graphics, vol. 38, no. 5, pp. 403–410, 2014.

[113] Y. Liu, R. Jin, and L. Yang, “Semi-supervised multi-label learning by constrained non-negative matrix factorization,” in Proceedings of the national conference on artificial intelligence, vol. 21, no. 1, 2006, p. 421.

[114] Z. Liu, X. Li, C. C. Luo, P.and Loy, and X. Tang, “Semantic image segmentation via deep parsing network,” in IEEE International Conference on Computer Vision (ICCV, December 2015, pp. 1377–1385.

[115] M. Loane, H. Gore, R. Corbett, K. Steele, C. Mathews, S. Bloomer, D. Eedy, R. Telford, and R. Wootton, “Effect of camera performance on diagnostic accuracy: preliminary results from the northern ireland arms of the uk multicentre teledermatology trial,” Journal of Telemedicine and Telecare, vol. 3, no. 2, pp. 83–88, 1997.

[116] D. Lowe, “Distinctive image features from scale-invariant keypoints,” International journal of computer vision, vol. 60, no. 2, pp. 91–110, 2004.

[117] V. K. Madasu and B. Lovell, “Blotch detection in pigmented skin lesions using fuzzy co-clustering and texture segmentation,” in Digital Image Computing: Techniques and Applications (DICTA). IEEE, 2009, pp. 25–31.

[118] A. Madooei and M. S. Drew, “Incorporating colour information for computer-aided diagnosis of melanoma from dermoscopy images: A retrospective survey and critical analysis,” International Journal of Biomedical Imaging, vol. 2016, 2016.

[119] A. Madooei, M. S. Drew, M. Sadeghi, and M. S. Atkins, “Automatic detection of blue-white veil by discrete colour matching in dermoscopy images,” in Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013. Springer, 2013, pp. 453–460.

[120] A. Madooei, M. Drew, and H. Hajimirsadeghi, “Learning to detect blue-white structures in dermoscopy images with weak supervision,” CoRR, vol. abs/1506.09179, 2015. [Online]. Available: http://arxiv.org/abs/1506.09179

[121] I. Maglogiannis and C. Doukas, “Overview of advanced computer vision systems for skin lesions characterization,” IEEE Transactions on Information Technology in Biomedicine, vol. 13, no. 5, pp. 721–733, 2009.

[122] J. Mairal, F. Bach, and J. Ponce, “Sparse modeling for image and vision processing,” Foundations and Trends in Computer Graphics and Vision, vol. 8, no. 2-3, pp. 85–283, 2014.

[123] A. R. S. Marcal, T. Mendonca, C. S. P. Silva, M. A. Pereira, and R. J., “Evaluation of the menzies method potential for automatic dermoscopic image analysis,” in Computational Modelling of Objects Represented in Images - CompImage 2012, 2012, pp. 103–108.

[124] A. Martinez-Uso, F. Pla, and J. M. Sotoca, “A semi-supervised gaussian mixture model for image segmentation,” in Pattern Recognition (ICPR), 2010 20th International Conference on, 2010, pp. 2941–2944.

[125] J. Mayer, “Systematic review of the diagnostic accuracy of dermatoscopy in detecting malignant melanoma.” The Medical Journal of Australia, vol. 167, no. 4, pp. 206–210, 1997.

[126] T. Mendonça, P. M. Ferreira, J. Marques, A. R. Marcal, and J. Rozeira, "Ph 2-a dermoscopic image database for research and benchmarking," in 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, 2013, pp. 5437–5440.

[127] C. Mendoza, C. Serrano, and B. Acha, "Scale invariant descriptors in pattern analysis of melanocytic lesions," in 16th IEEE International Conference on Image Processing (ICIP). IEEE, 2009, pp. 4193–4196.

[128] S. Merler, C. Furlanello, B. Larcher, and A. Sboner, "Tuning cost-sensitive boosting and its application to melanoma diagnosis," in Multiple classifier systems. Springer, 2001, pp. 32–42.

[129] M. Messadi, A. Bessaid, and A. Taleb-Ahmed, "Extraction of specific parameters for skin tumour classification," Journal of medical engineering & technology, vol. 33, no. 4, pp. 288–295, 2009.

[130] K. Mikolajczyk and C. Schmid, "Scale & affine invariant interest point detectors," International journal of computer vision, vol. 60, no. 1, pp. 63–86, 2004.

[131] H. Mirzaalian, T. Lee, and G. Hamarneh, "Learning features for streak detection in dermoscopic color images using localized radial flux of principal intensity curvature," in 2012 IEEE Workshop on Mathematical Methods in Biomedical Image Analysis (MMBIA). IEEE, 2012, pp. 97–101.

[132] Y. Mori, H. Takahashi, and R. Oka, "Image-to-word transformation based on dividing and vector quantizing images with words," in First International Workshop on Multimedia Intelligent Storage and Retrieval Management. Citeseer, 1999, pp. 1–9.

[133] A. Murali, W. Stoecker, and R. Moss, "Detection of solid pigment in dermatoscopy images using texture analysis," Skin Research and Technology, vol. 6, no. 4, pp. 193–198, 2000.

[134] J. C. Nascimento and J. S. Marques, "Performance evaluation of object detection algorithms for video surveillance," IEEE Transactions on multimedia, vol. 8, pp. 761–773, 2006.

[135] P. Natarajan, S. Wu, S. Vitaladevuni, X. Zhuang, S. Tsakalidis, U. Park, R. Prasad, and P. Natarajan, "Multimodal feature fusion for robust event detection in web videos," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR),. IEEE, 2012, pp. 1298–1305.

[136] M. H. Nguyen, L. Torresani, F. de la Torre, and C. Rother, "Weakly supervised discriminative localization and classification: a joint learning process," in IEEE 12th International Conference on Computer Vision (ICCV). IEEE, 2009, pp. 1925–1932.

[137] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in IEEE International Conference on Computer Vision, 2015, pp. 1520–1528.

[138] L. A. Nowak, M. Ogorzalek, and M. Pawlowski, "Pigmented network structure detection using semi-smart adaptive filters," in 2012 IEEE 6th International Conference on Systems Biology (ISB). IEEE, 2012, pp. 310–314.

[139] R. Oliveira, J. Papa, A. Pereira, and J. Tavares, "Computational methods for pigmented skin lesion classification in images: review and future trends," Neural Computing and Applications, pp. 1–24, 2016.

[140] M. Paulin, J. Revaud, Z. Harchaoui, F. Perronnin, and C. Schmid, "Transformation pursuit for image classification," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3646–3653.

[141] H. Pehamberger, A. Steiner, and K. Wolff, "In vivo epiluminescence microscopy of pigmented skin lesions. i. pattern analysis of pigmented skin lesions," Journal of the American Academy of Dermatology, vol. 17, no. 4, pp. 571–583, 1987.

[142] G. Pellacani, C. Grana, and S. Seidenari, "Automated description of colours in polarized-light surface microscopy images of melanocytic lesions," Melanoma Research, vol. 14, no. 2, pp. 125–130, 2004.

[143] D. Perednia and N. Brown, "Teledermatology: one application of telemedicine." Bulletin of the Medical Library Association, vol. 83, no. 1, p. 42, 1995.

[144] C. Pleiss, J. Risse, H. Biersack, and H. Bender, "Role of fdg-pet in the assessment of survival prognosis in melanoma," Cancer biotherapy & radiopharmaceuticals, vol. 22, no. 6, pp. 740–747, 2007.

[145] C. Poynton, Digital video and HD: Algorithms and Interfaces. Morgan Kaufman, 2012.

[146] G. J. Qi, X. S. Hua, Y. Rui, J. Tang, T. Mei, and H. J. Zhang, "Correlative multi-label video annotation," in Proceedings of the 15th international conference on Multimedia. ACM, 2007, pp. 17–26.

[147] J. Quintana, R. Garcia, and L. Neumann, "A novel method for color correction in epiluminescence microscopy," Computerized Medical Imaging and Graphics, vol. 35, pp. 646–652, 2011.

[148] M. M. Rahman and P. Bhattacharya, "An integrated and interactive decision support system for automated melanoma recognition of dermoscopic images," Computerized Medical Imaging and Graphics, vol. 34, no. 6, pp. 479–486, 2010.

[149] M. Rastgoo, R. Garcia, O. Morel, and F. Marzani, "Automatic differentiation of melanoma from dysplastic nevi," Computerized Medical Imaging and Graphics, vol. 43, pp. 44–52, 2015.

[150] J. Read, B. Pfahringer, G. Holmes, and E. Frank, "Classifier chains for multi-label classification," Machine learning, vol. 85, no. 3, pp. 333–359, 2011.

[151] P. Rubegni, G. Cevenini, M. Burroni, R. Perotti, G. Dell'Eva, P. Sbano, C. Miracco, P. Luzi, P. Tosi, and P. Barbini, "Automated diagnosis of pigmented skin lesions," International Journal of Cancer, vol. 101, no. 6, pp. 576–580, 2002.

[152] M. Ruela, C. Barata, J. S. Marques, and J. Rozeira, “A system for the detection of melanomas in dermoscopy images using shape and symmetry features,” Accepted for publication in Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, 2015.

[153] S. Rui, W. Jin, and T. Chua, “A novel approach to auto image annotation based on pairwise constrained clustering and semi-na ï ve bayesian model,” in Proceedings of the 11th International Multimedia Modelling Conference. IEEE, 2005, pp. 322–327.

[154] D. Ruiz, V. Berenguer, A. Soriano, and B. Sánchez, “A decision support system for the diagnosis of melanoma: A comparative approach,” Expert Systems with Applications, vol. 38, no. 12, pp. 15 217–15 223, 2011.

[155] M. Sadeghi, T. Lee, H. Lui, D. McLean, and S. Atkins, “Detection and analysis of irregular streaks in dermoscopic images of skin lesions,” IEEE Transactions on Medical Imaging, vol. 32, pp. 849–861, 2013.

[156] M. Sadeghi, T. K. Lee, D. McLean, H. Lui, and M. S. Atkins, “Global pattern analysis and classification of dermoscopic images using textons,” in SPIE Medical Imaging. International Society for Optics and Photonics, 2012, pp. 83 144X–83 144X.

[157] M. Sadeghi, T. Lee, D. McLean, H. Lui, and M. Atkins, “Oriented pattern analysis for streak detection in dermoscopy images,” in Medical Image Computing and Computer-Assisted Intervention–MICCAI. Springer, 2012, pp. 298–306.

[158] M. Sadeghi, M. Razmara, T. K. Lee, and M. S. Atkins, “A novel method for detection of pigment network in dermoscopic images using graphs,” Computerized Medical Imaging and Graphics, vol. 35, no. 2, pp. 137–143, 2011.

[159] M. Sadeghi, M. Razmara, P. Wighton, T. K. Lee, and M. S. Atkins, “Modeling the dermoscopic structure pigment network using a clinically inspired feature set,” in International Workshop on Medical Imaging and Virtual Reality. Springer, 2010, pp. 467–474.

[160] A. Sáez, C. Serrano, and B. Acha, “Model-based classification methods of global patterns in dermoscopic images,” IEEE Transactions on Medical Imaging, vol. 33, no. 5, pp. 1137–1147, 2014.

[161] A. Sboner, P. Bauer, G. Zumiani, C. Eccher, E. Blanzieri, S. Forti, and M. Cristofolini, “Clinical validation of an automated system for supporting the early diagnosis of melanoma,” Skin Research and Technology, vol. 10, no. 3, pp. 184–192, 2004.

[162] R. E. Schapire and Y. Singer, “Boostexter: A boosting-based system for text categorization,” Machine learning, vol. 39, no. 2, pp. 135–168, 2000.

[163] P. Schmid-Saugeon, J. Guillod, and J. Thiran, “Towards a computer-aided diagnosis system for pigmented skin lesions,” Computerized Medical Imaging and Graphics, vol. 27, no. 1, pp. 65–78, 2003.

[164] S. Seidenari, G. Pellacani, and C. Grana, “Computer description of colours in dermoscopic melanocytic lesion images reproducing clinical assessment,” British Journal of Dermatology, vol. 149, no. 3, pp. 523–529, 2003.

[165] ——, “Colors in atypical nevi: a computer description reproducing clinical assessment,” Skin Research and Technology, vol. 11, no. 1, pp. 36–41, 2005.

[166] ——, “Asymmetry in dermoscopic melanocytic lesion images: a computer description based on colour distribution,” Acta dermato-venereologica, vol. 86, no. 2, pp. 123–128, 2006.

[167] C. Serrano and B. Acha, “Pattern analysis of dermoscopic images based on markov random fields,” Pattern Recognition, vol. 42, no. 6, pp. 1052–1057, 2009.

[168] S. Shafer, “Using color to separate reflection components,” Color Research & Application, vol. 10, no. 4, pp. 210–218, 1985.

[169] D. Shen and H. Ip, “Discriminative wavelet shape descriptors for recognition of 2-d patterns,” Pattern recognition, vol. 32, no. 2, pp. 151–165, 1999.

[170] J. Shi and J. Malik, “Normalized cuts and image segmentation,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 8, pp. 888–905, 2000.

[171] K. Shimizu, H. Iyatomi, M. Celebi, K. Norton, and M. Tanaka, “Four-class classification of skin lesions with task decomposition strategy,” IEEE Transactions on Biomedical Engineering, vol. 62, no. 1, pp. 274–283, 2015.

[172] K. Shimizu, H. Iyatomi, K. A. Norton, and M. E. Celebi, “Extension of automated melanoma screening for non-melanocytic skin lesions,” International Journal of Computer Applications in Technology, vol. 50, no. 1-2, pp. 122–130, 2014.

[173] B. Shrestha, J. Bishop, K. Kam, X. Chen, R. Moss, W. Stoecker, S. Umbaugh, R. Stanley, M. Celebi, and A. Marghoob, “Detection of atypical texture features in early malignant melanoma,” Skin Research and Technology, vol. 16, no. 1, pp. 60–65, 2010.

[174] M. Silveira, J. Nascimento, J. S. Marques, A. Marçal, T. Mendonça, S. Yamauchi, J. Maeda, and J. Rozeira, “Comparison of segmentation methods for melanoma diagnosis in dermoscopy images,” IEEE Journal of Selected Topics in Signal Processing, vol. 3, no. 1, pp. 35–45, 2009.

[175] S. Singh, J. Stevenson, and D. McGurty, “An evaluation of polaroid photographic imaging for cutaneous-lesion referrals to an outpatient clinic: a pilot study,” British journal of plastic surgery, vol. 54, no. 2, pp. 140–143, 2001.

[176] J. Sivic and A. Zisserman, “Video google: A text retrieval approach to object matching in videos,” in Ninth IEEE International Conference on Computer Vision (ICCV). IEEE, 2003, pp. 1470–1477.

[177] C. Snoek, M. Worring, and A. Smeulders, “Early versus late fusion in semantic video analysis,” in 13th Annual ACM international conference on Multimedia, 2005, pp. 399–402.

[178] J. Solomon, S. Mavinkurve, D. Cox, and R. Summers, “Computer-assisted detection of subcutaneous melanomas: feasibility assessment,” Academic radiology, vol. 11, no. 6, pp. 678–685, 2004.

[179] S. Sra, “A short note on parameter approximation for von mises-fisher distributions: and a fast implementation of $I_s(x)$,” Computational Statistics, vol. 27, no. 1, pp. 177–190, 2012.

[180] W. Stoecker, K. Gupta, R. Stanley, R. Moss, and B. Shrestha, “Detection of asymmetric blotches (asymmetric structureless areas) in dermoscopy images of malignant melanoma using relative color,” Skin Research and Technology, vol. 11, no. 3, pp. 179–184, 2005.

[181] W. Stoecker, M. Wronkiewiecz, R. Chowdhury, R. Stanley, J. Xu, A. Bangert, B. Shrestha, D. Calcara, H. Rabinovitz, and M. Oliviero, “Detection of granularity in dermoscopy images of malignant melanoma using color and texture features,” Computerized Medical Imaging and Graphics, vol. 35, no. 2, pp. 144–147, 2011.

[182] W. Stolz, A. Riemann, and A. B. Cognetta, “ABCD rule of dermatoscopy: a new practical method for early recognition of malignant melanoma,” European Journal of Dermatology, vol. 4, pp. 521–527, 1994.

[183] T. Tanaka, S. Torii, I. Kabuta, K. Shimizu, and M. Tanaka, “Pattern classification of nevus with texture analysis,” IEEJ Transactions on Electrical and Electronic Engineering, vol. 3, no. 1, pp. 143–150, 2008.

[184] M. Tkalcic and J. F. Tasic, “Colour spaces: perceptual, historical and applicational background,” in Eurocon, vol. 1, 2003, pp. 304–308.

[185] A. Torralba, R. Fergus, and W. T. Freeman, “80 million tiny images: A large data set for nonparametric object and scene recognition,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 30, no. 11, pp. 1958–1970, 2008.

[186] J. Tran, “Segmentation of dermatological images using mixture models and markov random fields,” 2005.

[187] K. Van De Sande, T. Gevers, and C. Snoek, “Evaluating color descriptors for object and scene recognition,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, no. 9, pp. 1582–1596, 2010.

[188] J. van de Weijer, T. Gevers, and A. Gijsenij, “Edge-based color constancy,” IEEE Transactions on Image Processing, vol. 16, no. 9, pp. 2207–2214, 2007.

[189] H. L. Van Trees, Detection, estimation, and modulation theory. John Wiley & Sons, 2004.

[190] A. Vedaldi and S. Soatto, “Quick shift and kernel methods for mode seeking,” in Computer Vision–ECCV 2008. Springer, 2008, pp. 705–718.

[191] S. Vijayanarasimhan and K. Grauman, “What’s it going to cost you?: Predicting effort vs. informativeness for multi-label image annotations,” in IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2009, pp. 2262–2269.

[192] P. Viola and M. J. Jones, “Robust real-time face detection,” International journal of computer vision, vol. 57, no. 2, pp. 137–154, 2004.

[193] J. von Kries, “Influence of adaptation on the effects produced by luminous stimuli,” Sources of Color Vision, pp. 109–119, 1970.

[194] X. J. Wang, L. Zhang, F. Jing, and W. Y. Ma, “Annosearch: Image auto-annotation by search,” in IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), vol. 2. IEEE, 2006, pp. 1483–1490.

[195] P. Wighton, T. K. Lee, H. Lui, D. McLean, and M. S. Atkins, “Chromatic aberration correction: an enhancement to the calibration of low-cost digital dermoscopes,” Skin Research and Technology, vol. 17, pp. 339–347, 2011.

[196] P. Wighton, T. K. Lee, H. Lui, D. I. McLean, and M. S. Atkins, “Generalizing common tasks in automated skin lesion diagnosis,” IEEE Transactions on Information Technology in Biomedicine, vol. 15, no. 4, pp. 622–629, 2011.

[197] I. Wolf, J. Smolle, H. Soyer, and H. Kerl, “Sensitivity in the clinical diagnosis of malignant melanoma,” Melanoma research, vol. 8, no. 5, pp. 425–429, 1998.

[198] R. Yan, J. Tesic, and J. R. Smith, “Model-shared subspace boosting for multi-label classification,” in Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2007, pp. 834–843.

[199] C. Yang, M. Dong, and J. Hua, “Region-based image annotation using asymmetrical support vector machine-based multiple-instance learning,” in IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), vol. 2. IEEE, 2006, pp. 2057–2063.

[200] J. Yang, K. Yu, Y. Gong, and T. Huang, “Linear spatial pyramid matching using sparse coding for image classification,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2009, pp. 1794–1801.

[201] T. Yoshida, M. E. Celebi, G. Schaefer, and H. Iyatomi, “Simple and effective pre-processing for automated melanoma discrimination based on cytological findings,” in IEEE International Conference on Big Data (Big Data). IEEE, 2016, pp. 3439–3442.

[202] S. Yoshino, T. Tanaka, M. Tanaka, and H. Oka, “Application of morphology for detection of dots in tumor,” in SICE 2004 Annual Conference, vol. 1. IEEE, 2004, pp. 591–594.

[203] Z. Zha, X. Hua, T. Mei, J. Wang, G. Qi, and Z. Wang, “Joint multi-label multi-instance learning for image classification,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2008, pp. 1–8.

[204] C. Zhang, J. Liu, Q. Tian, C. Xu, H. Lu, and S. Ma, “Image classification by non-negative sparse coding, low-rank and sparse decomposition,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2011, pp. 1673–1680.

[205] D. Zhang, M. M. Islam, and G. Lu, “A review on automatic image annotation techniques,” Pattern Recognition, vol. 45, no. 1, pp. 346–362, 2012.

[206] M. L. Zhang and K. Zhang, “Multi-label learning by exploiting label dependency,” in Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2010, pp. 999–1008.

[207] M. L. Zhang and Z. H. Zhou, “Multilabel neural networks with applications to functional genomics and text categorization,” IEEE Transactions on Knowledge and Data Engineering, vol. 18, no. 10, pp. 1338–1351, 2006.

[208] ——, “Ml-knn: A lazy learning approach to multi-label learning,” Pattern recognition, vol. 40, no. 7, pp. 2038–2048, 2007.

[209] ——, “A review on multi-label learning algorithms,” IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 8, pp. 1819–1837, 2014.

[210] Z. H. Zhou, M. L. Zhang, S. J. Huang, and Y. F. Li, “Multi-instance multi-label learning,” Artificial Intelligence, vol. 176, no. 1, pp. 2291–2320, 2012.

[211] S. Zhu, X. Ji, W. Xu, and Y. Gong, “Multi-labelled classification using maximum entropy method,” in Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2005, pp. 274–281.

A

A CAD System Based on Global Features - Complete Results

This appendix shows all of the results obtained with the different configurations described in Chapter 4. Four classifiers are considered (AdaBoost, SVM, kNN, and random forests) and the hyperparameters of each classifier (see Table 4.2) are chosen by nested cross validation.

Table A.1: Classification results obtained using the AdaBoost algorithm for the PH^2 and reduced sets. * signals the results obtained with the reduced set. **Bold** highlights the best results.

Feature	**Descriptor**	SE	SP	S
Color	**RGB Histograms**	**88%**	**86%**	**0.128**
	HSV Histograms	82%	75%	0.208
	L*a*b* Histograms	90%	83%	0.142
	Opponent Histograms	91%	81%	0.136
	RGB Histograms	50%*	76%*	0.396*
	HSV Histograms	75%*	90%*	0.190*
	L*a*b* Histograms	85%*	86%*	0.146*
	Opponent Histograms	**90%***	**79%***	**0.144***
Texture	**Amplitude Histogram**	**84%**	**77%**	**0.188**
	Orientation Histogram	66%	60%	0.364
	Gabor Filters	92%	61%	0.204
	Amplitude Histogram	**75%***	**77%***	**0.242***
	Orientation Histogram	15%*	76%*	0.606*
	Gabor Filters	70%*	65%*	0.320*
Shape	**Simple Shape Descriptors**	**80%***	**58%***	**0.288***
	Hu Moments	90%*	33%*	0.328*
	Wavelets	70%*	69%*	0.304*
Symmetry	Shape Symmetry	65%*	50%*	0.410*
	Color Symmetry	80%*	80%*	0.200*
	Texture Symmetry	**90%***	**73%***	**0.168***

Table A.2: Classification results obtained using the SVM algorithm for the PH^2 and reduced sets. * signals the results obtained with the reduced set. **Bold** highlights the best results

Feature	**Descriptor**	SE	SP	S
Color	RGB Histograms	73%	84%	0.226
	HSV Histograms	81%	86%	0.170
	L*a*b* Histograms	**87%**	**86%**	**0.134**
	Opponent Histograms	81%	77%	0.206
	RGB Histograms	65%*	81%*	0.286*
	HSV Histograms	**80%***	**80%***	**0.200***
	L*a*b* Histograms	75%*	77%*	0.242*
	Opponent Histograms	75%*	76%*	0.246*
Texture	Amplitude Histogram	74%	75%	0.256
	Orientation Histogram	62%	75%	0.328
	Gabor Filters	**86%**	**73%**	**0.192**
	Amplitude Histogram	55%*	71%*	0.386*
	Orientation Histogram	40%*	65%*	0.500*
	Gabor Filters	**75%***	**56%***	**0.326***
Shape	**Simple Shape Descriptors**	**85%***	**55%***	**0.270***
	Hu Moments	70%*	65%*	0.312*
	Wavelets	50%*	60%*	0.460*
Symmetry	Shape Symmetry	60%*	32%*	0.512*
	Color Symmetry	60%*	80%*	0.320*
	Texture Symmetry	**70%***	**68%***	**0.308***

Table A.3: Classification results obtained using the kNN algorithm for the PH^2 and reduced sets. * signals the results obtained with the reduced set. **Bold** highlights the best results

Feature	Descriptor	SE	SP	S
Color	RGB Histograms	84%	85%	0.156
	HSV Histograms	**89%**	**85%**	**0.126**
	L*a*b* Histograms	87%	82%	0.150
	Opponent Histograms	83%	78%	0.190
	RGB Histograms	70%*	69%*	0.304*
	HSV Histograms	90%*	72%*	0.172*
	L*a*b* Histograms	**100%***	**67%***	**0.132***
	Opponent Histograms	80%*	74%*	0.224*
Texture	**Amplitude Histogram**	**83%**	**84%**	**0.166**
	Orientation Histogram	55%	65%	0.410
	Gabor Filters	84%	71%	0.212
	Amplitude Histogram	**95%***	**84%***	**0.094***
	Orientation Histogram	50%*	66%*	0.436*
	Gabor Filters	65%*	67%*	0.342*
Shape	Simple Shape Descriptors	85%*	41%*	0.348*
	Hu Moments	80%*	43%*	0.370*
	Wavelets	**65%***	**60%***	**0.370***
Symmetry	Shape Symmetry	70%*	51%*	0.488*
	Color Symmetry	**80%***	**75%***	**0.220***
	Texture Symmetry	75%*	69%*	0.274*

Table A.4: Classification results obtained using the random forests algorithm for the PH^2 and reduced sets. * signals the results obtained with the reduced set. **Bold** highlights the best results

Feature	**Descriptor**	SE	SP	S
Color	RGB Histograms	88%	84%	0.136
	HSV Histograms	87%	86%	0.134
	L*a*b* Histograms	86%	86%	0.140
	Opponent Histograms	**89%**	**84%**	**0.130**
	RGB Histograms	50%*	75%*	0.400*
	HSV Histograms	**90%***	**86%***	**0.116***
	L*a*b* Histograms	75%*	84%*	0.214*
	Opponent Histograms	80%*	86%*	0.176*
Texture	**Amplitude Histogram**	**84%**	**80%**	**0.176**
	Orientation Histogram	61%	63%	0.382
	Gabor Filters	90%	63%	0.208
	Amplitude Histogram	**75%***	**73%***	**0.258***
	Orientation Histogram	50%*	56%*	0.478*
	Gabor Filters	75%*	58%*	0.318*
Shape	Simple Shape Descriptors	90%*	48%*	0.268*
	Hu Moments	**70%***	**81%***	**0.256***
	Wavelets	55%*	59%*	0.434*
Symmetry	Shape Symmetry	50%*	53%*	0.488*
	Color Symmetry	**90%***	**75%***	**0.160***
	Texture Symmetry	65%*	74%*	0.314*

B

The Role of Local Features - Complete Results

This appendix shows all of the results obtained with the different configurations described in Chapter 5. Four classifiers are considered (AdaBoost, SVM, kNN, and random forests) and the hyperparameters of each classifier (see Table 4.2) are chosen by nested cross validation.

Table B.1: Classification results obtained using the AdaBoost algorithm. **Bold** highlights the best results.

Feature	**Descriptor**	SE	SP	S
Color	RGB Histograms	85%	84%	0.154
	HSV Histograms	82%	89%	0.152
	L*a*b* Histograms	90%	82%	0.132
	Opponent Histograms	**90%**	**84%**	**0.124**
Texture	**Amplitude Histogram**	**87%**	**81%**	**0.154**
	Orientation Histogram	81%	82%	0.186
	Gabor Filters	88%	77%	0.164

Table B.2: Classification results obtained using the SVM algorithm. **Bold** highlights the best results

Feature	**Descriptor**	SE	SP	S
Color	RGB Histograms	87%	78%	0.166
	HSV Histograms	84%	82%	0.168
	L*a*b* Histograms	**87%**	**79%**	**0.162**
	Opponent Histograms	87%	75%	0.178
Texture	Amplitude Histogram	76%	89%	0.193
	Orientation Histogram	**78%**	**88%**	**0.182**
	Gabor Filters	62%	89%	0.273

Table B.3: Classification results obtained using the kNN algorithm. **Bold** highlights the best results

Feature	**Descriptor**	SE	SP	S
Color	**RGB Histograms**	**100%**	**76%**	**0.096**
	HSV Histograms	97%	75%	0.118
	L*a*b* Histograms	98%	79%	0.150
	Opponent Histograms	95%	78%	0.118
Texture	Amplitude Histogram	94%	70%	0.156
	Orientation Histogram	96%	63%	0.172
	Gabor Filters	**91%**	**77%**	**0.146**

Table B.4: Classification results obtained using the Random Forests algorithm. **Bold** highlights the best results

Feature	**Descriptor**	SE	SP	S
Color	RGB Histograms	94%	68%	0.164
	HSV Histograms	92%	79%	0.132
	L*a*b* Histograms	**94%**	**77%**	**0.128**
	Opponent Histograms	92%	78%	0.136
Texture	Amplitude Histogram	87%	85%	0.138
	Orientation Histogram	90%	79%	0.144
	Gabor Filters	**88%**	**88%**	**0.120**

C

corr-LDA Equations

C.1 Expanded Lower Bound $\mathcal{L}(\gamma^d, \phi^d, \lambda^d; \alpha, \beta, \Omega)$

This section shows the expanded version of every term in (9.12), such that it is possible to obtain a lower bound $\mathcal{L}(\gamma^d, \phi^d, \lambda^d; \alpha, \beta, \Omega)$ as a function of the model (α, β, Ω) and variational $(\gamma^d, \phi^d, \lambda^d)$ parameters. For details on these derivations please refer to [25, 26].

$$
\begin{aligned}
\mathcal{L}(\gamma^d, \phi^d, \lambda^d; \alpha, \beta, \Omega) \quad &= \quad \mathbb{E}_q \left[\log p(\theta^d|\alpha) . \left(\prod_{n=1}^{N^d} p(z_n^d|\theta^d) p(r_n^d|z_n^d, \Omega) \right) . \left(\prod_{m=1}^{M^d} p(y_m^d|N^d) p(w_m^d|y_m^d, \mathbf{z},^d \beta) \right) \right] \\
&- \quad \mathbb{E}_q \left[\log q(\theta^d|\gamma^d) . \left(\prod_{n=1}^{N^d} q(z_n^d|\phi_n^d) \right) \; . \; \left(\prod_{m=1}^{M^d} q(y_m^d|\lambda_m^d) \right) \right] , \qquad \text{(C.1)}
\end{aligned}
$$

which can be decomposed in

$$
\begin{aligned}
\mathcal{L}\left(\gamma^d, \phi^d, \lambda^d; \alpha, \beta, \Omega\right) \quad &= \quad \mathbb{E}_q[\log p(\theta^d|\alpha)] + \mathbb{E}_q[\log p(\mathbf{z}^d|\theta^d)] \\
&+ \quad \mathbb{E}_q[\log p(\mathbf{r}^d|\mathbf{z}^d, \Omega)] + \mathbb{E}_q[\log p(\mathbf{y}^d|N^d)] \\
&+ \quad \mathbb{E}_q[\log p(\mathbf{w}^d|\mathbf{y}^d, \mathbf{z}^d, \beta)] - \mathbb{E}_q[\log q(\theta^d|\gamma^d)] \\
&- \quad \mathbb{E}_q[\log q(\mathbf{z}^d|\phi^d)] - \mathbb{E}_q[\log q(\mathbf{y}^d|\lambda^d)] . \qquad \text{(C.2)}
\end{aligned}
$$

Each of the terms in (C.2) corresponds to the following expressions

$$\begin{aligned}\mathbb{E}_q[\log p(\theta^d|\alpha)] &= \log\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)-\sum_{k=1}^{K}\log\Gamma(\alpha_k)\\ &+ \sum_{k=1}^{K}(\alpha_k-1)\left(\Psi(\gamma_k^d)-\Psi\left(\sum_{j=1}^{K}\gamma_j^d\right)\right), \qquad \text{(C.3)}\end{aligned}$$

where $\Gamma(.)$ is the gamma function and $\Psi(.)$ is the digamma function, i.e., the first derivative of the gamma function.

$$\mathbb{E}_q[\log p(\mathbf{z}^d|\theta^d)] = \sum_{n=1}^{N^d}\sum_{k=1}^{K}\left(\Psi(\gamma_k^d)-\Psi\left(\sum_{j=1}^{K}\gamma_j^d\right)\right)\phi_{nk}^d. \qquad \text{(C.4)}$$

$$\mathbb{E}_q[\log p(\mathbf{r}^d|\mathbf{z}^d,\Omega)] = \sum_{n=1}^{N^d}\sum_{k=1}^{K}\phi_{nk}^d\log p(r_n^d|z_n^d=k,\Omega). \qquad \text{(C.5)}$$

$$\mathbb{E}_q[\log p(\mathbf{y}^d|N^d)] = C, \qquad \text{(C.6)}$$

where C is a constant.

$$\mathbb{E}_q[\log p(\mathbf{w}^d|\mathbf{y}^d,\mathbf{z}^d,\beta)] = \sum_{m=1}^{M^d}\sum_{n=1}^{N^d}\sum_{k=1}^{K}\lambda_{mn}^d\phi_{nk}^d\log p(w_m^d|y_m^d=n,z_n^d=k,\beta). \qquad \text{(C.7)}$$

$$\begin{aligned}\mathbb{E}_q[\log q(\theta^d|\gamma^d)] &= \log\Gamma\left(\sum_{k=1}^{K}\gamma_k^d\right)-\sum_{k=1}^{K}\log\Gamma(\gamma_k^d)\\ &+ \sum_{k=1}^{K}(\alpha_k-1)\left(\Psi(\gamma_k^d)-\Psi\left(\sum_{j=1}^{K}\gamma_j^d\right)\right). \qquad \text{(C.8)}\end{aligned}$$

$$\mathbb{E}_q[\log q(\mathbf{z}^d|\phi^d)] = \sum_{n=1}^{N^d}\sum_{k=1}^{K}\phi_{nk}^d\log\phi_{nk}^d. \qquad \text{(C.9)}$$

$$\mathbb{E}_q[\log q(\mathbf{y}^d|\lambda^d)] = \sum_{m=1}^{M^d}\sum_{n=1}^{N^d}\lambda_{mn}^d\log\lambda_{mn}^d. \qquad \text{(C.10)}$$

C.2 Variational EM - Update Equations

•**E-Step** The update equations for the variational parameters $(\gamma^d,\phi^d,\lambda^d)$ are the following

$$\begin{aligned}\phi_{nk}^d &\propto p(r_n^d|z_n=k,\Omega_{z_n})\exp\left\{E_q[\log q(\theta_k|\gamma^d)]\right\}.\\ &.\exp\left\{\sum_{m=1}^{M}\lambda_{mn}^d\log p(w_m^d|y_m=n,z_{y_m}=k,\beta_{z_{y_m}})\right\}\end{aligned} \qquad \text{(C.11)}$$

$$\lambda_{mn}^d \propto \exp\left\{\sum_{k=1}^{K}\phi_{nk}^d\log p(w_m^d|y_m=n,z_{y_m}=k,\beta_{z_{y_m}})\right\}, \qquad \text{(C.12)}$$

$$\gamma_k^d = \alpha_k + \sum_{n=1}^{N_d}\phi_{nk}^d. \qquad \text{(C.13)}$$

These parameters must be estimated by the order that they are presented here.

•**M-Step** The parameter β that relates the text labels w_m with the topic k is updated as follows

$$\beta_{km} \propto \sum_{d=1}^{D} w_m^d \sum_{n=1}^{N_d} \phi_{nk}^d \lambda_{mn}^d . \tag{C.14}$$

It is not possible to obtain an exact update equation for the Dirichlet parameter α . Therefore, Blei and Jordan propose the use of the Newton-Raphon's method [26] to obtain an estimate of this parameter.

When the traditional formulation of corr-LDA is used, each of the k multivariate parameters $\Omega_k = (\mu_k, \Sigma_k)$ is computed as follows:

$$\mu_k = \frac{\sum_{d=1}^{D} \sum_{n=1}^{N_d} \phi_{nk}^d r_n^d}{\sum_{d=1}^{D} \sum_{n=1}^{N_d} \phi_{nk}^d}, \tag{C.15}$$

$$\Sigma_k = \frac{\sum_{d=1}^{D} \sum_{n=1}^{N_d} \phi_{nk}^d (r_n^d - \mu_k)(r_n^d - \mu_k)^T}{\sum_{d=1}^{D} \sum_{n=1}^{N_d} \phi_{nk}^d}. \tag{C.16}$$

When the von Mises-Gaussian distributions are used to model the regions' features the update equations for $\Omega_k = (\mu_k, \Sigma_k, \tau_k, \varepsilon_k)$ are as follows. The parameters μ and Σ are update as in (C.15) and (C.16), but using r_n' (feature vector without the H channel information). The parameters of the von Mises distribution are updated using the following equations

$$\tau_k = \tan^{-1}\left(\frac{\sum_{d=1}^{D} \sum_{n=1}^{N_d} \phi_{nk}^d \sin \mathsf{H}_n^d}{\sum_{d=1}^{D} \sum_{n=1}^{N_d} \phi_{nk}^d \cos \mathsf{H}_n^d}\right), \tag{C.17}$$

An analytical computation of the parameter ε_k is not possible. Different approximations have been proposed to tackle this issue. This thesis adopts the approach described in [179], which makes use of the Newton-Raphson's method to obtain an approximation. This method requires a few iterations t of the following equation:

$$\varepsilon_k^t = \varepsilon_k^{t-1} - \frac{A(\varepsilon_k^{t-1}) - \overline{R}}{1 - A(\varepsilon_k^{t-1})}, \tag{C.18}$$

where,

$$A(\varepsilon_k^{t-1}) = \frac{I_1(\varepsilon_k^{t-1})}{I_0(\varepsilon_k^{t-1})}, \tag{C.19}$$

and the variable $\overline{R}$ is defined as

$$\overline{R} = \frac{\sum_{d=1}^{D} \sum_{n=1}^{N_d} \phi_{nk}^d \cos([\mathsf{H}]_n^d - \tau_k)}{\sum_{d=1}^{D} \sum_{n=1}^{N_d} \phi_{nk}^d}. \tag{C.20}$$

In the first iteration ε_k^0 is set to be [179]

$$\varepsilon_k^0 = \frac{\overline{R} - \overline{R}^3}{1 - \overline{R}^2}. \tag{C.21}$$

The update equation is applied until convergence is reached.

D

Towards the Development of a Clinically Inspired System - Complete Results

This appendix shows all of the results obtained in Chapter 9. Two classifiers are considered (SVM with RBF kernel and random forests), as well as two fusion strategies (LR and MR). The hyperparameters of each classifier are chosen by nested cross validation.

Table D.1: Results for the detection of texture structures: pigment network and dots. **Bold** highlights the best results.

Features	**Structures**	**SVM-RBF**	**Random Forests**
μ_c, μ_{ca}	Pigment Network	$Re = 67.1\%$ $Pre = 69.7\%$ $F1 = 68.4\%$	$Re = 72.6\%$ $Pre = 71.3\%$ $F1 = 71.9\%$
	Dots/Globules	$Re = 70.7\%$ $Pre = 66.7\%$ $F1 = 68.6\%$	$Re = 68.3\%$ $Pre = 67.9\%$ $F1 = 68.1\%$
μ_M, σ_M, μ_m, σ_m	Pigment Network	$Re = 83.8\%$ $Pre = 79.4\%$ $F1 = 81.5\%$	$Re = 86.3\%$ $Pre = 77.4\%$ $F1 = 81.6\%$
	Dots/Globules	$Re = 72.8\%$ $Pre = 71.0\%$ $F1 = 71.9\%$	$Re = 76.5\%$ $Pre = 69.8\%$ $F1 = 73.0\%$
μ_c, μ_{ca}, μ_M, σ_M, μ_m, σ_m	Pigment Network	$Re = 82.6\%$ $Pre = 78.4\%$ $F1 = 80.4\%$	$Re = 80.9\%$ $Pre = 74.9\%$ $F1 = 77.3\%$
	Dots/Globules	$Re = 71.5\%$ $Pre = 70.8\%$ $F1 = 71.1\%$	$Re = 66.7\%$ $Pre = 69.1\%$ $F1 = 67.9\%$
μ_c, μ_{ca}, μ_M, σ_M, μ_m, σ_m, μ_g, σ_g	Pigment Network	$Re =$ **82.6**$\%$ $Pre =$ **80.0%** $F1 =$ **81.3%**	$Re =$ **88.9%** $Pre =$ **77.6**$\%$ $F1 =$ **82.9**$\%$
	Dots/Globules	$Re =$ **74.8**$\%$ $Pre =$ **71.3**$\%$ $F1 =$ **73.0**$\%$	$Re =$ **83.2**$\%$ $Pre =$ **71.8**$\%$ $F1 =$ **77.1**$\%$

Table D.2: Results for the detection of color structures: blue-whitish veil and regression areas. **Bold** highlights the best results.

Features	Structures	SVM-RBF	Random Forests
μ_{HSV},μ_c, μ_{ca}	Blue-whitish veil	$Re =$ **72.6%** $Pre =$ **71.8**$\%$ $F1 =$ **72.1**$\%$	$Re =$ **68.5%** $Pre =$ **75.4**$\%$ $F1 =$ **71.9**$\%$
	Regression areas	$Re =$ **47.8**$\%$ $Pre =$ **55.3**$\%$ $F1 =$ **51.3**$\%$	$Re =$ **51.3**$\%$ $Pre =$ **60.8**$\%$ $F1 =$ **55.6**$\%$
μ_{HSV}, μ_M,σ_M, μ_m, σ_m	Blue-whitish veil	$Re = 67.1\%$ $Pre = 81.0\%$ $F1 = 73.4\%$	$Re = 66.4\%$ $Pre = 78.7\%$ $F1 = 72.0\%$
	Regression areas	$Re = 42.9\%$ $Pre = 57.2\%$ $F1 = 49.0\%$	$Re = 28.4\%$ $Pre = 70.5\%$ $F1 = 40.5\%$
μ_c, μ_{ca}, μ_M, σ_M, μ_m, σ_m, μ_{HSV}	Blue-whitish veil	$Re = 70.3\%$ $Pre = 72.0\%$ $F1 = 71.0\%$	$Re = 67.9\%$ $Pre = 77.2\%$ $F1 = 72.3\%$
	Regression areas	$Re = 38.9\%$ $Pre = 50.0\%$ $F1 = 43.6\%$	$Re = 40.5\%$ $Pre = 51.2\%$ $F1 = 45.2\%$

Table D.3: Results for melanoma diagnosis using combinations of criteria. All* refers to the combination of the colors' model with the one associated with the color and texture structures.

	SVM-RBF		Random Forests		Combined	
Criteria	LR	MR	LR	MR	LR	MR
Colors + Texture Structures	$SE = 77.5\%$ $SP = 67.3\%$ $AUC = 80.0\%$	$SE = 78.8\%$ $SP = 68.4\%$ $AUC = 82.7\%$	$SE = 74.7\%$ $SP = 79.2\%$ $AUC = 83.1\%$	$SE = 78.8\%$ $SP = 75.1\%$ $AUC = 83.8\%$	$SE = 77.2\%$ $SP = 75.1\%$ $AUC = 82.3\%$	$SE = 81.3\%$ $SP = 73.5\%$ $AUC = 85.3\%$
Colors + Color Structures	$SE = 77.9\%$ $SP = 69.1\%$ $AUC = 81.1\%$	$SE = 78.8\%$ $SP = 72.8\%$ $AUC = 82.5\%$	$SE = 82.1\%$ $SP = 71.8\%$ $AUC = 83.6\%$	$SE = 81.3\%$ $SP = 71.2\%$ $AUC = 83.7\%$	$SE = 79.6\%$ $SP = 72.5\%$ $AUC = 84.1\%$	$SE = 81.3\%$ $SP = 72.5\%$ $AUC = 84.9\%$
Texture + Color Structures	$SE = 84.6\%$ $SP = 66.8\%$ $AUC = 83.6\%$	$SE = 85.9\%$ $SP = 67.1\%$ $AUC = 84.4\%$	$SE = 73.9\%$ $SP = 78.9\%$ $AUC = 82.6\%$	$SE = 78.8\%$ $SP = 73.0\%$ $AUC = 84.7\%$	$SE = 79.2\%$ $SP = 76.0\%$ $AUC = 82.7\%$	$SE = 85.9\%$ $SP = 72.8\%$ $AUC = 86.3\%$
All	$SE = 81.3\%$ $SP = 67.0\%$ $AUC = 82.5\%$	$SE = 84.2\%$ $SP = 72.1\%$ $AUC = 85.1\%$	$SE = 73.4\%$ $SP = 78.2\%$ $AUC = 84.0\%$	$SE = 81.3\%$ $SP = 74.8\%$ $AUC = 85.4\%$	$SE = 78.0\%$ $SP = 77.1\%$ $AUC = 85.6\%$	$SE =$ **84.6%** $SP =$ **74.2%** $AUC =$ **86.9%**
All*	$SE = 82.5\%$ $SP = 69.1\%$ $AUC = 83.2\%$	$SE = 85.4\%$ $SP = 65.4\%$ $AUC = 83.8\%$	$SE = 78.8\%$ $SP = 76.7\%$ $AUC = 83.8\%$	$SE = 80.9\%$ $SP = 73.2\%$ $AUC = 84.3\%$	$SE = 79.2\%$ $SP = 75.6\%$ $AUC = 83.6\%$	$SE =$ **85.8%** $SP =$ **71.1%** $AUC =$ **85.9%**