Michel Silva Fornaciali

Towards Robust Melanoma Screening: A Case for Enhanced Mid-level Features

Triagem Robusta de Melanoma: Em Defesa dos Descritores Aprimorados de Nível Médio

Campinas

2015

UNIVERSIDADE ESTADUAL DE CAMPINAS
Faculdade de Engenharia Elétrica e de Computação

Michel Silva Fornaciali

Towards Robust Melanoma Screening: A Case for Enhanced Mid-level Features

Triagem Robusta de Melanoma: Em Defesa dos Descritores Aprimorados de Nível Médio

Supervisor: Prof. Dr. Eduardo Alves do Valle Junior

Co-supervisor: Dr[a]. Sandra Eliza Fontes de Avila

Master's dissertation presented to the Post Graduate Program of the School of Electrical and Computer Engineering of the University of Campinas to obtain a Master's degree in Electrical Engineering, in the area of concentration Computer Engineering.

This volume corresponds to the version of the dissertation submitted to the examining board by Michel Silva Fornaciali, under the supervision of Prof. Dr. Eduardo Alves do Valle Junior and Dr[a]. Sandra Eliza Fontes de Avila

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Engenharia Elétrica da Faculdade de Engenharia Elétrica e de Computação da Universidade Estadual de Campinas para obtenção do título de Mestre em Engenharia Elétrica, na área de concentração Engenharia de Computação.

Este exemplar corresponde à versão da dissertação apresentada à banca examinadora pelo aluno Michel Silva Fornaciali, sob orientação de Prof. Dr. Eduardo Alves do Valle Junior e Dr[a]. Sandra Eliza Fontes de Avila

---

Campinas

2015

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca da Área de Engenharia e Arquitetura
Elizangela Aparecida dos Santos Souza - CRB 8/8098

F727t

Fornaciali, Michel Silva, 1988-
Towards robust melanoma screening : a case for enhanced mid-level features / Michel Silva Fornaciali. – Campinas, SP : [s.n.], 2015.

Orientador: Eduardo Alves do Valle Junior.
Coorientador: Sandra Eliza Fontes de Avila.
Dissertação (mestrado) – Universidade Estadual de Campinas, Faculdade de Engenharia Elétrica e de Computação.

1. Melanoma. 2. Imagem. 3. Classificação. I. Valle Junior, Eduardo Alves do,1978-. II. Avila, Sandra Eliza Fontes de. III. Universidade Estadual de Campinas. Faculdade de Engenharia Elétrica e de Computação. IV. Título.

Informações para Biblioteca Digital

Título em outro idioma: Triagem robusta de melanoma : em defesa dos descritores aprimorados de nível médio
Palavras-chave em inglês:
Melanoma
Image
Classification
Área de concentração: Engenharia de Computação
Titulação: Mestre em Engenharia Elétrica
Banca examinadora:
Eduardo Alves do Valle Junior [Orientador]
Flávia Vasques Bittencourt
Roberto de Alencar Lotufo
Data de defesa: 25-06-2015
Programa de Pós-Graduação: Engenharia Elétrica

iv

COMISSÃO JULGADORA - TESE DE MESTRADO

Candidato: Michel Silva Fornaciali

Data da Defesa: 25 de junho de 2015

Título da Tese: "Towards Robust Melanoma Screening: A Case for Enhanced Mid-level Features (Triagem Robusta de Melanoma: Em Defesa dos Descritores Aprimorados de Nível Médio)"

Prof. Dr. Eduardo Alves do Valle Junior (Presidente): Eduardo Valle

Profa. Dra. Flávia Vasques Bittencourt: Flavia B Hexsourt

Prof. Dr. Roberto de Alencar Lotufo:

Abstract

Melanoma is the skin cancer that most leads to death, even while being the most curable when detected early. Melanoma diagnosis, however, is a difficult task, requiring special training. This poses a challenge for poor and isolated communities, where the full-time presence of a specialist is unfeasible. Therefore, automated screening appears as an attractive solution, allowing to refer to the doctor only the patients at higher risk. Much of the existing art on automated melanoma screening is based on the Bag-of-Visual-Words (BoVW) model, combining color and texture descriptors. However, the BoVW model has been improving and nowadays there are several extensions that deliver better classification rates. Those enhanced models have not yet been explored for melanoma screening, thus motivating our work. Here we present a new approach for melanoma screening, based upon the state-of-the-art BossaNova descriptors, showing very promising results, reaching an AUC of up to 93.7%. This work also proposes a new spatial pooling strategy specially designed for melanoma screening.

Keywords: Melanoma, Dermoscopy, Automated Screening, Image Classification, Mid-level Features.

Resumo

Melanoma é o câncer de pele que mais leva à morte, mesmo sendo o mais curável quando detectado precocemente. O diagnóstico do melanoma, no entanto, é uma tarefa difícil, exigindo treinamento especial. Isto representa um desafio para comunidades pobres e isoladas nas quais a presença em tempo integral de um especialista é inviável. Assim, o rastreio automático aparece como uma solução atrativa, permitindo encaminhamento médico apenas para os pacientes com alto risco. Muitos trabalhos existentes sobre rastreio automático de melanoma são baseados no modelo de Bag-of-Visual-Words (BoVW), combinando descritores de cor e textura. No entanto, o modelo BoVW tem se aprimorado e hoje em dia existem várias extensões que oferecem melhores taxas de classificação. Estes modelos avançados ainda não foram explorados para a triagem do melanoma, motivando assim nosso trabalho. Aqui nós apresentamos uma nova abordagem para rastreio do melanoma, baseado nos descritores BossaNova, que são estado-da-arte, mostrando resultados muitos promissores, atingindo uma AUC de até 93,7%. Este trabalho também propõe uma nova estratégia de *pooling* espacial especialmente projetada para o rastreio do melanoma.

Palavras-chaves: Melanoma, Dermoscopia, Triagem Automática, Classificação de Imagens, Descritores de Nível Médio.

Contents

Acknowledgements

First and foremost, I would like to thank Prof. Eduardo Valle for being a great mentor and advisor. His commitment was crucial to the success of this work and if now I can consider myself a researcher is due to all the teachings transmitted in the time we were together. He was always present in the highs and lows of the work always pushing me to my best. I also must thank my co-advisor Sandra Avila for her priceless advices and attention during all the research, especially in experimental designs and critical analysis of practical data. I also thank my colleagues from our research group for all valuable discussions and knowledge sharing.

I am very grateful of my parents (Silvia and Daniel) who always supported me and encouraged me to run for my dreams. Since I was a child they taught me that Education is the basis of success. I also thank my friends and my dear boyfriend Felipe Reis for understanding the stresses and absences in social life due to this work.

I would like to thank Eldorado Research Institute for supporting this dissertation by motivating my studies and understanding my absences in critical times of the research.

Everybody knows that it is impossible to get a Msc without resources. So, I would like to thank UNICAMP/FEEC for the administration support, CAPES/FAPESP for the financial support and RECOD Lab / CENAPAD / Microsoft Azure / Amazon for the infrastructure used in the experiments.

At last, but not least, I would like to thank Madonna for her songs that rocked several nights of experiments and writing of papers and, of course, this dissertation.

Once again, thank you very much for everything! This work is a result of all of us.

Now, let's do all again in my PhD!

List of Figures

List of Tables

List of Acronyms and Abbreviations

ANN Artificial Neural Networks

ANOVA Analysis of Variance

BoVW Bag-of-Visual-Words

BoVW Bag-of-Visual-Words

BoW Bag-of-Words

BoW Bag-of-Words

CVPR Conference on Computer Vision and Pattern Recognition

DLA Deep Learning Architectures

DoG Difference-of-Gaussian

GMM Gaussian Mixture Model

HoG Histogram of Gradient

LASC Locality-Constrained Affine Subspace Coding

LLC Locality-Constrained Linear Coding

PCA Principal Component Analysis

SCP Spatial Circular Pooling

SIFT Scale-Invariant Feature Transform

SPM Spatial Pyramid Matching

SURF Speeded Up Robust Features

SVC Super-Vector Coding

VLAD Vector of Locally Aggregated Descriptors

VLAT Vector of Locally Aggregated Tensors

1 Introduction

Computer-aided diagnosis has growing importance in medicine, empowering doctors with tools for decision making. Among the different medical data that can be analyzed by computers, medical images deserves especial mention, due to the recent impressive advances of Computer Vision.

In this work, we are particularly interested in **automated screening**. Screening, in medicine, is a strategy for identifying a latent disease in individuals, who may not necessarily present obvious signs or symptoms. Screening allows finding illness as early as possible, facilitating the treatment, improving the prognosis, and, in case of severe diseases, reducing the risk of serious lesions, and even death. The World Health Organization published in 1968 the Wilson's criteria, a set of rules that must be obeyed for a screening program to be successful. Among them, we highlight that *facilities for diagnosis and treatment should be available* and *case-finding should be a continued process* (instead of a one-shot procedure). By reducing costs and increasing availability, automated screening improves the odds that a screening program will succeed, especially where the permanent presence of a medical specialist is not economically feasible (e.g., in rural, isolated, or poor communities).

Different branches of medicine can benefit from automated screening through images: cardiology (echocardiography), neurology (Alzheimer), oncology (mammography), ophthalmology (diabetic retinopathy) and so on [Abedini et al., 2015; Neltner et al., 2012; Skaane et al., 2013; Faust et al., 2012]. In this work, however, our focus is on Dermatology, specifically the screening of melanoma, being the type of skin cancer that most leads to death, but curable if detected early [SCF, 2013; ACS, 2013].

1.1 Motivation

Melanoma is the type of skin cancer that most leads to death if treatment is delayed, because of its malignancy (frequent occurrence of metastases) [SCF, 2013; ACS, 2013]. Nevertheless, it is a curable cancer if detected early. This reinforces the need of effective screening strategies for melanoma, particularly, again, in communities where the continuous presence of a dermatologist is not feasible.

Melanoma is also the type of cancer whose incidence most increased: according to Rigel [2010], the risk of an American developing invasive malignant melanoma was 1 in 1,500 in the 1930s. In the 2010s, this number jumped to 1 in 59.

This research also has high potential to be exploited in other scenarios. As smartphones get improved camera quality, and increased computing power, they become attractive for automated screening tasks as well [Bastawrous, 2012]. Thus, the choice of this theme opens opportunity for future investigations of mobile "smart" devices able to provide the automated screening.

1.2 Problem Statement

In the point of view of Computer Vision, screening by images is clearly an image classification task. Image classification consists in using the visual content of an image to determine the category to which it belongs. It is a Machine Learning task, in which first a model is estimated from a set of annotated images (training set), and then the model is used to predict the class of other images. That process is represented on Figure 1.

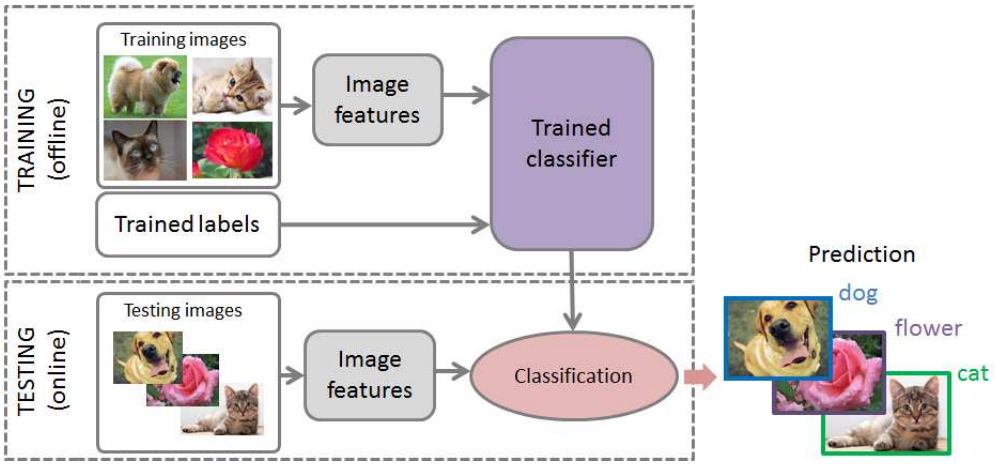

Figure 1 – Classical representation of an image classification system.

While we have studied and reviewed the broader topic of image classification for many types of datasets, we have chosen to focus our original contributions on medical images, and more specifically, in the screening of melanoma from dermoscopic or clinical images. Therefore, the problem to be addressed by this research can be formulated as:

> *Given an image of a skin lesion, how to automatically classify it into melanoma or not?*

Medical-imaging classification is a type of special-purpose image classification, that is, the classification of images belonging to datasets with few types of classes. The so called

"specific datasets" have limited semantical scope, for example, images of flowers, birds, or cats. The limited scope of special-purpose image classification is, on one hand, an advantage, since the researcher can optimize the processing for a specific task. On the other hand, in general-purpose classification — that aims at a large number of classes, with great semantic variability — we usually are interested in broad categories (e.g., flower or bird) that are easier to identify than the fine-grained and often subtle categorization of special-purpose classification (e.g., identifying specific species of flowers or birds).

1.3 Challenges

Research in automated screening and computer-aided diagnosis face several challenges. The most serious is perhaps the scarcity of large-enough, publicly-available datasets of annotated images, which are essential for both training and validating the classification models. Often each research group has its own private dataset, which is not available, not even by request. This makes the direct comparison of techniques proposed by different groups very challenging, hurting the possibility of effective meta-analysis. Moreover, high accuracy is needed for those applications, lest they are not really useful for improving the screening process. Existing end-user apps for melanoma screening are too inaccurate, as was pointed by Wolf et al. [2013], and might mislead patients to a false sense of security, making them forgo a medical appointment. Those low-accurate solutions often lead doctors to distrust automatic methods altogether.

Although the objectives of a research like ours are shared by both medical and computing researchers, those communities often work alone with little mutual cooperation. This is another factor that explains the challenge in obtaining good quality data for training processes; but it also explains the difficulty in analyzing and interpreting the empirical results.

1.4 Objectives

Our main goal is to validate a modern approach for automated melanoma screening. As a second objective, we aim to investigate image classification techniques that can benefit other problems related to "specific datasets".

Specific objectives of this research are:

- Improving the accuracy of automated melanoma screening.
- Advancing the global understanding of the problem by providing a critical review of works already present in literature.

- Opening the opportunity for investigations of melanoma screening in mobile environment using new approaches.
- Opening the opportunity for extending our screening techniques for other diseases.
- Opening the opportunity to advance the understanding of models for special-purpose image classification.

1.5 Contributions

In turn, our main contributions are:

- Novel techniques for melanoma screening.
- A comparative survey of melanoma screening techniques.
- A protocol for experimentation on melanoma screening that help to make studies reproducible and comparable.

1.6 Outline

The reminder of the text is organized as follows.

Chapter 2 – Literature Review We establish the foundations of this work by describing the related works available in literature about image classification in the last sixteen years. This chapter also shows the main studies about melanoma screening and discuss its major aspects.

Chapter 3 – Proposed Solution We present our approach to melanoma screening contrasting it with methods present in literature. We also discuss the main open issues in literature that are addressed by this work.

Chapter 4 – Experimental Results We validate our hypotheses through empirical data collected in several experiments. Each experiment is described in terms of its central hypothesis, experimental design, results and analysis.

Chapter 5 – Conclusions We discuss the impact of our research in the literature of melanoma screening, present our concluding remarks, propose future work directions, and also explore possible extensions for this research.

2 Literature Review

The growing power of Image Processing and Computer Vision explains their increasing adoption in Medicine. Several types of images, ranging from 2D to 4D, from conventional X-rays to real-time tomography, can now be analyzed automatically or semi-automatically, with increasing accuracies.

Computing power can be exploited to extract information not easily perceived by humans, expanding the power of doctors and facilitating the diagnosis and treatment of serious diseases. Just to illustrate a few examples, we can cite the work of Rondina et al. [2002] for segmenting cardiac magnetic resonance images, as well as the works of Pires et al. [2014b,a], for screening diabetic retinopathy. We also highlight the use of magnetic resonance imaging in neurology [Castellano et al., 2003], for segmenting brain structures [Rittner et al., 2009], and also classifying regions of interest, and types of brain lesions [Bento et al., 2013].

Furthermore, computational power may allow large-scale screening programs, decreasing distances between patients and doctors, accelerating the diagnosis, and lowering costs. Such topic has worldwide relevance, especially for its applications to underserved communities.

This work deals with the screening of diseases by images, specifically the case of melanoma. Successfully screening any disease means distinguishing the healthy patient from the sick one. In this particular case, we aim to determine whether or not a skin lesion, given its image, is a melanoma. Our tools will be Computer Vision, Pattern Recognition, and Image Classification.

To facilitate reading, we divided this chapter in three sections. First, we start discussing about general-purpose image classification, and how it is currently addressed (Section 2.1). After that, we present the traditional BoVW for image classification (Section 2.2). In the following section, we narrow our focus to our key application, reviewing the existing art on automated melanoma screening (Section 2.3). Concluding this chapter, Section 4.4 summarizes the main information showed here and also opens the discussions of our proposed solution.

2.1 Image Classification

The internet, together with the availability of cheap image-capturing devices, have created an explosion of visual content. This, by itself, has motivated a pressing interest on the automatic

classification of videos and images.

When compared to text retrieval/classification, image classification is even more challenging, because the content, in the form of pixels has very little direct meaning, in opposition to the words and sentences of text. This very large "semantic gap" of visual information is a recurrent theme of research in Information Retrieval, Image Processing, and Computer Vision [Smeulders et al., 2000; Lowe, 2004; Perronnin et al., 2010; Avila et al., 2013].

Nowadays there are several approaches to image classification. Although they are algorithmically distinct, all of them rely on the same typical concepts: (i) feature extraction from the pixels, (ii) a robust description of the previous features, and (iii) a supervised classification. The main current techniques for image classification are the traditional BoVW and the ANN models.

The BoVW model was proposed by Sivic and Zisserman [2003] and also exploited by Csurka et al. [2004]. The metaphor was inspired from the BoW model from Textual Information Retrieval [Baeza-Yates and Ribeiro-Neto, 1999], where a document is represented by the frequency of words, without regard to higher-level structures (e.g., phrases). The classical BoVW model describes an image as a histogram of the occurrence rate of "visual words" in a "visual vocabulary" (or codebook) induced by quantifying the space of local image features, without attention to higher-level image organization (e.g., position of the features in the image).

In turn, ANN is older than BoVW. They were first proposed in 1943 by McCulloch and Pitts [1943]. ANN are inspired in biological models that try to simulate the existing neurons connections in our brain and how they interchange information, that is, ANN tries to recreate in computers how we, humans, learn. Mathematically, ANN are based in statistical learning algorithms that involves huge numbers of neurons organized in a net as inputs, hidden points and outputs. Each connection is weighted and its numeric value can change based in experience, enabling neural networks to adapt itself to different kinds of inputs and, therefore, being able to learn.

The BoVW model and its recent extensions are among the most used techniques for image classification. Nevertheless, DLA, as Deep Neural Networks [Krizhevsky et al., 2012] and Deep Belief Networks [Hinton, 2009] have recently appeared as the most competitive alternative for pattern recognition in images. DLA are an extension of ANN. They employ multiple layers of nonlinear processing units (making them "deep") and also different supervised or unsupervised learning mechanisms in each of these layers (making them very powerful to learn information from rough data).

Although a complete analytical understand of both BoVW and DLA for image clas-

sification is still lacking, it is known that both solutions have complementary advantages and issues. BoVW models are less flexible than DLA, but also much less greedy in terms of computing resources and annotated data. On the other hand, DLA suffer from the need to estimate huge numbers of parameters, implying the need of large training sets and a lot of computational resources.

Our current focus is on BoVW models, since they offer good accuracy without the need of extensive amounts of annotated data. This is critical for specific datasets especially for medical applications. This model will be fully explained in the next section since it is the basis of our solution.

2.2 Image Classification Through the BoVW Model

Among the current techniques for image classification, the BoVW model is one of the most studied approaches in literature. There are several implementations of it, each one with its own particularities, but in general the methods are based on the process detailed in Figure 2.

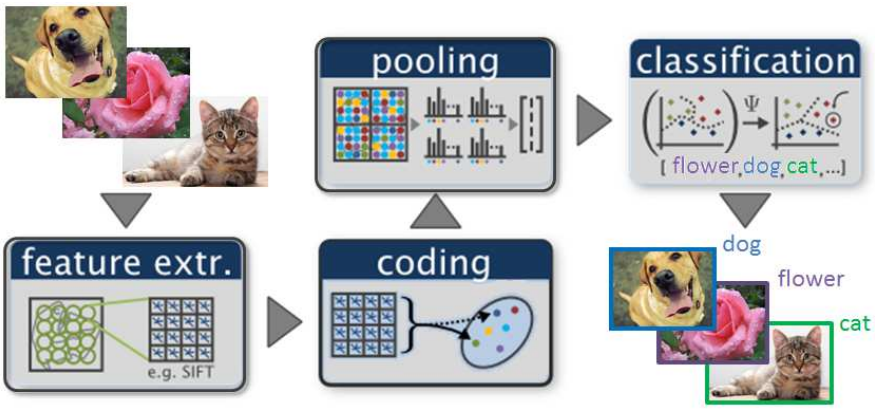

Figure 2 – Main pipeline of the BoVW model. The feature extraction is the low-level stage. The mid-level is decomposed in coding and pooling and the classification is generally done by a supervised method. Figure adapted from Chatfield et al. [2011]

The pipeline can be decomposed into three stages: **low-level feature extraction** (which extracts information directly from the image pixels), **mid-level feature extraction** (which makes the representation more general, aggregating abstraction to the model), and **classification** (a machine learning technique allowing the extraction of a general model from the individual data presented). The mid-level representation is the core of the BoVW proposition, and will be discussed more in-depth in the next sections.

The first stage, **low-level feature extraction**, consists of detecting and extracting local descriptors of the image. Local descriptors are visual features that represent small

patches of the image (in contrast to global features, which represent the entire image at once). The most cited descriptors are SIFT [Lowe, 2004] and SURF [Bay et al., 2006]. SIFT was created by Lowe in 2004, and is composed of a DoG interest region detector, followed by a HoG feature descriptor. The final vector has 128 dimensions and is invariant to uniform scaling, orientation, and partially invariant to affine distortion and illumination changes. SURF was introduced by Bay et al. [2006]. Inspired by the SIFT descriptor, SURF is several times faster than SIFT, since is based on sums of 2D Haar wavelet transforms and makes an efficient use of integral images. If the reader needs a complete comparison of invariant interest point detectors, we recommend the survey by Mikolajczyk and Schmid [2005] in which they compare the performance of descriptors computed for local interest regions. We also recommend the survey from Tuytelaars and Mikolajczyk [2008]. In that survey they define the properties of the ideal local feature detector and give a literature review over the past four decades. Other recommendation is the survey by Li and Allinson [2008], in which they provide a brief introduction for new researchers to the local feature research field, in order to facilitate the choice of an appropriate methodology according to specific requirements.

The last stage of the pipeline is the **classification** itself. As a machine learning task, it can be done *ad-hoc* with any technique desired. The main types of classification are supervised and unsupervised algorithms.

In supervised learning, the predictive model is constructed based on a set of examples called training set, that is a amount of data composed by the image and its label. The machine learning algorithm should be able to abstract the training set and construct a generalization based on it, being capable to determine the label of a new image presented to the predictive model. Meanwhile, in unsupervised learning, the training set is not labeled, and the algorithm itself is responsible to detect similarities within the data.

Classification in BoVW models is usually done with supervised methods, especially Support Vector Machines [Vapnik, 1995].

2.2.1 Mid-level Features: the Key Point of the BoVW

The low-level features are excessively discriminant, that is, very powerful to match the exactly same object or scene, but weak to identify categories or classes. For classification it is essential to improve the abstracting power of the model. That motivates the second stage of image classification through BoVW: the **mid-level feature extraction**. Essentially, the mid-level feature extraction is a powerful abstraction to the low-level features, that is, the low-level features are rewritten into a new space quantified by a codebook.

So, in the training phase, mid-level feature extraction must be preceded by **visual**

codebook learning, often accomplished with unsupervised learning, e.g., using k-means clustering to find a set of representative centroids, or an Expectation-Maximization procedure to estimate a GMM. Often, however, it is sufficient to just select at random a number of features from the training set to use as codewords.

Mid-level was formalized by Boureau et al. [2010] as the application of two successive steps: *coding* and *pooling*. The coding step transform the low-level features into a new representation based upon the codebook, and the pooling takes the average of the encoded features over the entire image. Since the pooling operation compacts all the information contained in the individually encoded local descriptors into a single feature vector, that step is critical for BoVW-based representations. In general terms, the objective of pooling is to summarize the information contained in the individually encoded descriptors into a single feature vector, preserving important information while discarding irrelevant detail [Avila et al., 2013].

The classical BoVW model employs *hard assignment* for the coding, and *averaging* for the pooling. Hard assignment associates each feature vector to the closest codeword. The final feature vector is obtained by averaging the encoded features. For a deeper comparison of mid-level feature coding and pooling approaches, we recommend the survey produced by Koniusz et al. [2013]. This traditional BoVW approach has important limitations, and several alternatives to that standard scheme have been recently developed. For instance, to attenuate the effect of coding errors induced by the descriptor-space quantization, hard quantization can be replaced by a soft assignment [van Gemert et al., 2010] or by other coding strategies such as sparse coding [Boureau et al., 2010]. Pooling by taking the maximum value (max-pooling) often performs better than average-pooling.

2.2.2 BoVW Formalism

After a practical description of the BoVW model, we discuss a more a formal definition of this technique. Although the formalism described here refers to the early BoVW approaches, with coding and pooling operations proposed by Boureau et al. [2010], the matrix notation is newer and was proposed by Benois-Pineau et al. [2012, chap. 3, section 3.1.1].

As mentioned before, the key aspect of the BoVW model is to describe each image in a notation that can be directly compared, instead of pixel values that are meaningless. This notation is done based in a codebook and occurs in the mid-level feature extraction.

First of all, in the low-level feature extraction stage the image is described as a numerical representation of its areas of interest (the so called *patches*). This is done, for example, with SIFT or SURF descriptors. The result of this stage are feature vectors with N dimensions. For SIFT, N is equal to 128. This process is briefly described in Figure 3.

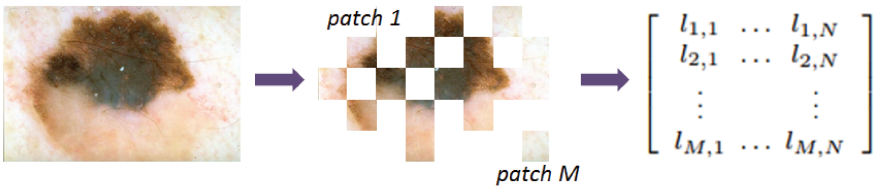

Figure 3 – Low-level feature extraction. The image is decomposed in special patches (small fragments with interest information) and each patch is described as a feature vector with fixed size N.

The codebook is nothing more than a set of low-level descriptors that represents the features of all images in the training set. It can be constructed using a clustering algorithm (to guarantee that the descriptors selected are, indeed, the most representative ones of the entire set) or, surprisingly, just picking random feature vectors of the available set. A key parameter of the codebook is its size, that is, the number of selected feature vectors it contains. These selected feature vectors are called *codewords*. In the following examples, let's consider that the codebook has M codewords.

After constructing the codebook, it is time to rewrite the low-level features into a new and more general representation. This representation is based in the codebook. Let's consider a matrix $\mathbf{H}$ with $M \times N$ size in which the lines represent each codeword of the codebook and the columns represent each low-level feature vector of the image. Figure 4 illustrates the $\mathbf{H}$ matrix.

The $\mathbf{H}$ matrix has $M \times N$ values. Each value, called α, is obtained during the *coding* step showed in Figure 5. It is calculated the distance of each j vector for each codeword i. Thus the $\alpha_{i,j}$ is 1 if this is the pair with the lowest distance (closest codeword from this feature vector) or 0 otherwise. Since both features and codewords are vectors, the distances can be calculated using, for example, Euclidean distance. This is the most traditional coding schema and is called *hard assignment*. In this case, it is easy to note that $\mathbf{H}$ is a sparse matrix.

In the *pooling* step, the information of the $\mathbf{H}$ matrix is summarized into a single vector $\mathbf{z}$ of size M. The procedure is illustrated in Figure 6. The most traditional pooling schema is averaging the values of each line of the $\mathbf{H}$ matrix. This leads to the final mid-level representation $\mathbf{z}$ of an image. This new representation is, therefore, based in and has the same size of the codebook.

$$\mathbf{H} = \begin{array}{c} \\ \mathbf{c}_1 \\ \vdots \\ \mathbf{c}_m \\ \vdots \\ \mathbf{c}_M \end{array} \begin{array}{c} \begin{array}{ccccccc} \mathbf{x}_1 & \dots & \mathbf{x}_j & \dots & \mathbf{x}_N \end{array} \\ \begin{bmatrix} \alpha_{1,1} & \dots & \alpha_{1,j} & \dots & \alpha_{1,N} \\ \vdots & & \vdots & & \vdots \\ \alpha_{m,1} & \dots & \alpha_{m,j} & \dots & \alpha_{m,N} \\ \vdots & & \vdots & & \vdots \\ \alpha_{M,1} & \dots & \alpha_{M,j} & \dots & \alpha_{M,N} \end{bmatrix} \end{array}$$

Figure 4 – Matrix $\mathbf{H}$ representing the relationship between M-codewords and N-feature vectors. Figure reproduced from Benois-Pineau et al. [2012].

$$\mathbf{H} = \begin{array}{c} \\ \mathbf{c}_1 \\ \vdots \\ \mathbf{c}_m \\ \vdots \\ \mathbf{c}_M \end{array} \begin{array}{c} \begin{array}{ccccccc} \mathbf{x}_1 & \dots & \mathbf{x}_j & \dots & \mathbf{x}_N \end{array} \\ \begin{bmatrix} \alpha_{1,1} & \dots & \alpha_{1,j} & \dots & \alpha_{1,N} \\ \vdots & & \vdots & & \vdots \\ \alpha_{m,1} & \dots & \alpha_{m,j} & \dots & \alpha_{m,N} \\ \vdots & & \vdots & & \vdots \\ \alpha_{M,1} & \dots & \alpha_{M,j} & \dots & \alpha_{M,N} \end{bmatrix} \end{array}$$

$$\textbf{Coding: } \mathbf{x}_j \rightarrow f(\mathbf{x}_j) = \{\alpha_{m,j}\}, \quad \alpha_{m,j} = 1 \text{ iff } m = \underset{k \in \{1,\dots,M\}}{\arg\min} \|\mathbf{x}_j - \mathbf{c}_k\|_2^2$$

Figure 5 – The coding step of the mid-level feature extraction. This is the case for hard assignment. Figure adapted from Avila [2013].

$$\mathbf{H} = \begin{array}{c} \\ \mathbf{c}_1 \\ \vdots \\ \mathbf{c}_m \\ \vdots \\ \mathbf{c}_M \end{array} \begin{array}{c} \begin{array}{ccccccc} \mathbf{x}_1 & \dots & \mathbf{x}_j & \dots & \mathbf{x}_N \end{array} \\ \begin{bmatrix} \alpha_{1,1} & \dots & \alpha_{1,j} & \dots & \alpha_{1,N} \\ \vdots & & \vdots & & \vdots \\ \alpha_{m,1} & \dots & \alpha_{m,j} & \dots & \alpha_{m,N} \\ \vdots & & \vdots & & \vdots \\ \alpha_{M,1} & \dots & \alpha_{M,j} & \dots & \alpha_{M,N} \end{bmatrix} \end{array} \qquad \mathbf{z} = \begin{bmatrix} z_1 \\ \vdots \\ z_m \\ \vdots \\ z_M \end{bmatrix}$$

$$\textbf{Pooling: } g(\{\alpha_j\}) = \mathbf{z} : \ \forall m, \ z_m = \sum_{j=1} \alpha_{m,j}$$

$$\textbf{BoVW representation: } \mathbf{z} = [z_1, z_2, \cdots, z_M]^\mathsf{T}$$

Figure 6 – The pooling step of the mid-level feature extraction. This is the case for average pooling. Figure adapted from Avila [2013].

2.2.3 Enhanced Mid-level Representations

From the last section, it is easy to note that the traditional BoVW can ignore important details of the images decreasing the representation power of the model. Nowadays there are many BoVW extensions that enrich the standard model, by preserving more information about the image. Among the cutting-edge BoVW representations, the most relevant are SPM [Lazebnik et al., 2006], Fisher Vector [Perronnin and Dance, 2007], [Perronnin et al., 2010], SVC [Zhou et al., 2010], VLAD [Jégou et al., 2010], VLAT [Picard and Gosselin, 2011], BossaNova [Avila et al., 2013] and LASC [Li et al., 2015]. Except for SPM and BossaNova, which keeps the representation compact, all of these approaches result in very large feature vectors, up to hundreds of thousands of dimensions.

The coding-pooling strategy enables image description abstracting the details of the level of the pixels. Nevertheless it doesn't take into account the spatial distribution of the elements along the image, that is, it is not possible to determine if a specific color or texture is *close* to another. The spatial information is very important for us, humans, understand and interpret images. To overcome the loss of spatial pooling information, Lazebnik et al. [2006] inaugurated the modern trend on BoVW approaches by proposing the SPM. Although this technique was first designed for recognizing scene categories, now it is used for improving several image classification problems and also explored in Neural Networks systems [He et al., 2014; Akata et al., 2014]. SPM splits an image into hierarchical regions, generating independent feature vectors that are concatenated to create the final representation. The feature vector of each region corresponds to the pooling of the encoded features vectors contained that region. Figure 7 illustrates how feature descriptors are quantified by the SPM technique.

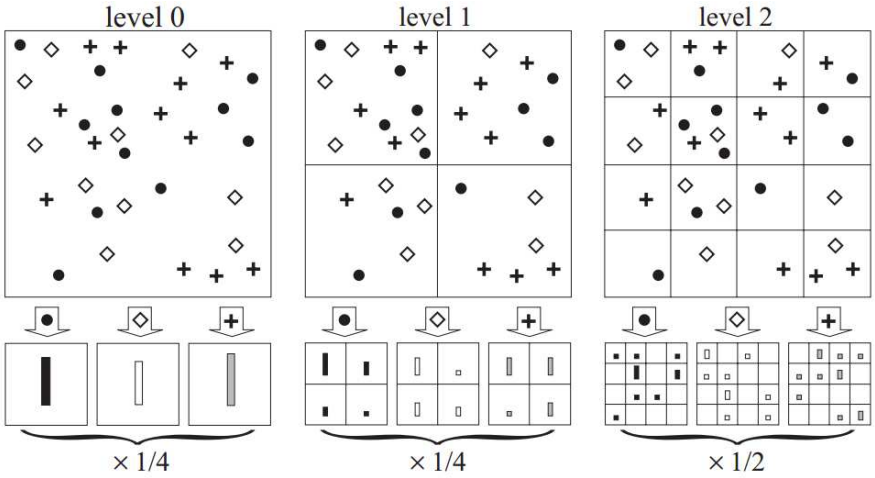

Figure 7 – An illustration of a three-level pyramid constructed by Spatial Pyramid Matching. Supposing that the image has three feature types (circles, diamonds and crosses), the image is subdivided in three different levels of resolution. For each level, the features are counted accordingly to the spatial bin they fell. Figure reproduced from Lazebnik et al. [2006]

.

Another BoVW extension is the Fisher Vector framework proposed bye Perronnin and Dance [2007]. The framework combines the strengths of generative and discriminative approaches. The idea is to characterize a signal with a gradient vector derived from a generative model and then pass this vector to a discriminative classifier. For image classification, the images are the input signals and the generative model is the codebook constructed by a GMM. The framework was also improved for large-scale image classification, applying a two-step normalization and spatial pyramids [Perronnin et al., 2010].

Still in 2010, we had two more extensions to the BoVW model: SVC [Zhou et al., 2010] and VLAD [Jégou et al., 2010]. SVC follows the same *coding-pooling-classification* schema, but they made contributions in the three steps. The coding phase is a simple extension of vector quantization coding that achieves a lower function approximation error. The pooling step is based on a novel probability kernel incorporating the similarity metric of local descriptors. On the other hand, VLAD can be seen as a simplification of the Fisher kernel. The idea is to accumulate for each codeword from the codebook, the differences of each local descriptor assigned to it. This characterizes the distribution of the vectors with respect to the center. VLAD was also improved by Arandjelovic and Zisserman [2013], generating MultiVLAD, a multiple spatial VLAD representation enabling retrieval and localization of objects that only extend over a small part of an image.

In 2011, VLAD was extended by Picard and Gosselin [2011] generating the VLAT model. Their final descriptor is composed by two types of elements: the first is the same of

VLAD (the sum of differences between the vectors "local descriptor" and "cluster center" associated with it); the second type is the sum of outer product of the same vectors.

Among the BoVW new approaches, we also highlight BossaNova [Avila et al., 2013]. BossaNova is a mid-level image representation which offers a better information-preserving pooling operation based on a distance-to-codeword distribution. It will be discussed in more details in Chapter 3 – Proposed Solution.

Although the DLA are gaining huge attention in recent researches for image classification, the mid-level representations also continue to evolve. Most of the recent art are new coding and/or pooling schemes. Since the coding is the critical step of the mid-level feature extraction, a comparative comprehensive study of the current literature is desired. A survey in this sense was recently proposed by Huang et al. [2014]. They discuss the feature coding methods in terms of motivations and mathematical representations.

Many of those new approaches, as well as Lazebnik et al. [2006], try to take advantages of the spatial distribution of the features along the image. Regardless the successful of the SPM, the technique requires nonlinear classifiers to achieve good image classification performance. One approach to overcome this issue is the LLC proposed by Wang et al. [2010]. Unlike SPM, LLC uses locality constraints to project each descriptor into its local-coordinate system, and them applies max pooling to produce the final representation. The final mid-level representation works well with linear classifiers, even with very large codebooks, enabling fast processing.

Thanks to its efficiency, the LLC method is suitable for many scenarios of image classification. Nevertheless, it discards the geometry of the feature space since each feature is projected in a simpler local space. One extension to attenuate this problem is the LASC proposed by Li et al. [2015]. In this approach, the feature vector is a composition of the top-k neighboring subspaces in which the descriptor is linearly decomposed, preserving, thus, more information about the geometry around it.

Therefore, we conclude that the image classification methods will continue to develop, mainly thanks to technological advances that allow complex processing in less time. Moreover, the mixture of BoVW and DLA models is a tendency, since they present complementary advantages [Li et al., 2014; Klein et al., 2015].

2.3 Melanoma Screening

Melanoma screening is an important matter on medical community, and it justifies the amount of researches on this field, explained by the increase of incidence cases among the

population [Rigel, 2010]. This section describes the most relevant researches of automated melanoma screening on the last seven years, based upon computer vision techniques. Table 1 summarizes the main information described here. Because there is no standard dataset neither protocol to allow direct comparison, the results reported by literature are not directly comparable. Note as well, that some authors employ AUC while others employ the accuracy (ACC) as metric.

Computer vision researchers tend to use *dermoscopic images* (those captured by specific medical devices — a dermatoscope, Figure 8 — in controlled conditions of acquisition, enabling better visualization of the lesion), instead of *clinical images* (captured by a common camera under non-controlled conditions) to classify skin lesions automatically due to the better quality, generally highlighting the lesion and its color and texture structures. Figure 9 illustrates the differences between dermoscopic and clinical images.

Figure 8 – A dermoscopy kit highlighting a dermatoscope: the instrument used by a physician to analyse skin lesions. Figure reproduced from the Internet.

Most of these studies tries to reproduce in computer machines the steps that dermatologists use to diagnose a melanoma. To accomplish it, some researches [Iyatomi et al., 2008; Mete and Sirakov, 2012; Abbas et al., 2012; Capdehourat et al., 2011] implement the ABCD Rule of Dermoscopy [Nachbar et al., 1994]. Others, for example Wadhawan et al. [2011b], employ the 7-Points Checklist [Argenziano et al., 1998] to classify a skin lesion.

The ABCD rule, also known as ABCDE rule, is a simple checklist for clinical diagnosis of melanoma. This rule looks for *asymmetry*, *border* shape, *color* aspects, *diameter* of the lesion and if it is *evolving*. Figure 10 illustrates how these aspects are analyzed by a physician. According to the ABCDE rule, the malignant lesion is asymmetrical. If you draw a line through its middle, the two halves will not match, indicating a sign of melanoma (Figure 10-(a)). Unlike melanomas, benign lesions have smooth borders. The borders of a melanoma

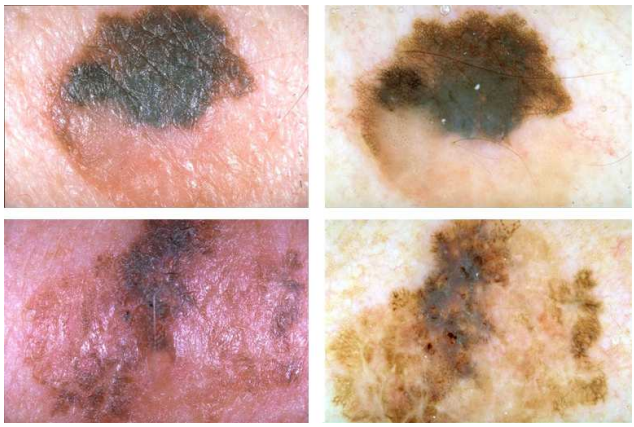

Figure 9 – A comparison between clinical (left) and dermoscopic (right) images. Example images from [Argenziano et al., 2002].

lesion tend to be uneven (Figure 10-(b)). Another relevant aspect is the color of the lesion: benign ones present usually a single color. Having different colors in the same lesion is a warning signal of melanoma, which may present shades of red, white or blue (Figure 10-(c)). Melanomas usually have a bigger diameter than benign lesions, but sometimes are smaller when they are detected (Figure 10-(d)). Finally, the last aspect is the lesion's evolution: in adults, benign lesions have the same size over time. If a lesion starts to change its size, shape, color or any other morphological aspect, it may be a serious warning of melanoma (Figure 10-(e)).

The 7-points checklist was proposed by Argenziano et al. [1998]. According to its authors, the checklist is an additional diagnostic algorithm developed for simplifying the classic pattern analysis of a skin lesion proposed in the Consensus Meeting of 1990. The main benefits of the new approach are the low number of features to identify and a scoring system to support reliable diagnostics. The 7-points are organized in two groups according to their probability to indicate a melanoma occurrence. The major criteria are (i) typical pigment network, (ii) blue-whitish veil and (iii) atypical vascular pattern. The minor criteria are (iv) irregular streaks, (v) irregular pigmentation, (vi) irregular dots/globules and (vii) regression structures. Each major criteria has score of two. The minor ones have score of one. The final score of a lesion is just the sum of scores for each criteria presented. If the final score is equal or greater than three, the lesion is a melanoma. Otherwise, it is a benign lesion. To exemplify how the "7-Points Checklist" is applied, Figure 11 shows the analysis of two lesions. The first, (a), has a final score of seven, which indicates that it is a melanoma. The second, (b), has a final score of one, which indicates that is is a benign lesion.

Any case have challenges that must be overcome, such as soft borders, which turn

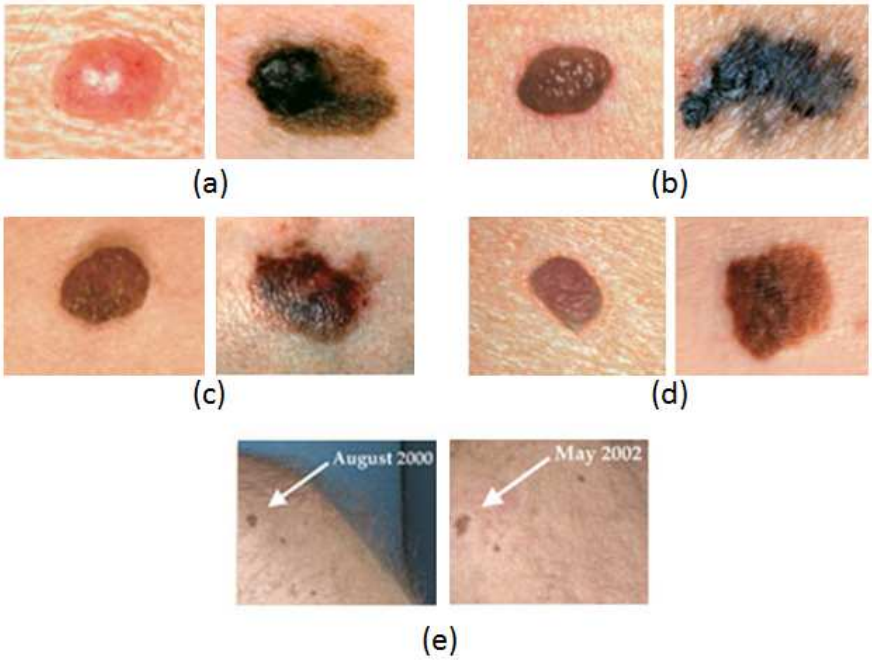

Figure 10 – Examples of skin lesion classification according to the ABCDE rule. For each pair, the image on right is malignant and the one on left is benign (unless for (e), which both are two pictures of the same benign lesion). (a) is related to asymmetry, (b) to border, (c) to color, (d) for diameter and (e) for how the lesion is evolving. Figure adapted from SCF [2013].

border detection into a hard problem, and the presence of veins or hair, which can impact the quality of the classification. For example, Abbas et al. [2012] deals with hair removal using derivative of Gaussian, morphological function, and fast marching techniques.

The ABCD Rule and the 7-Points Checklist evolved to modern approaches like the 3-Points Checklist [Soyer et al., 2004] and the 7-Points Checklist Revisited [Argenziano et al., 2011]. The 3-Points Checklist was designed as a simplification of the 7-Points in order to improve the reproducibility and the validity of the dermoscopy done be non-experts, which was proved through practical evaluations. The 7-Points Revisited is an evaluation of the diagnostic performance of pattern analysis with a lower threshold for excision.

2.3.1 Current Methods

According to the literature, the process of analyzing an image of a skin lesion has three main steps: (i) identify the lesion borders (*border detection*), (ii) extract image features only inside the lesion (*feature extraction*), and (iii) compare these features with pre-calculated features of both melanoma and non-melanoma examples to decide if the skin lesion is a melanoma or not (*classification*).

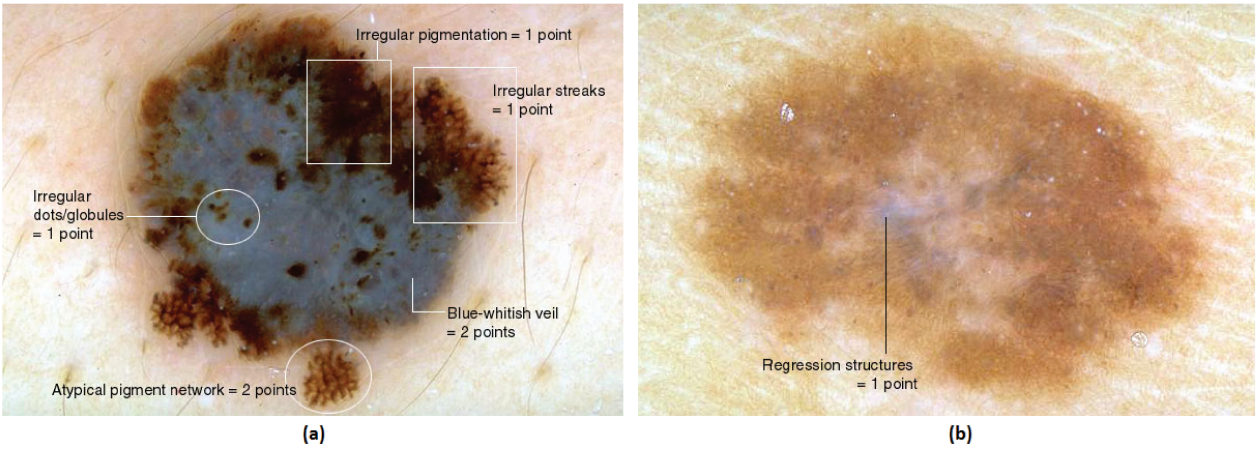

Figure 11 – Examples of dermoscopic lesion analysis by the 7-points checklist. It is attributed a point for each dermoscopic criteria present in the lesion, according to its severity. The points are summed leading to a final score of 7 (for lesion (a)) and 1 for lesion (b). This indicates that (a) is a melanoma and (b) is a benign lesion. Figures reproduced from [Argenziano et al., 1998].

Border detection can be found on Wadhawan et al. [2011a], which implements three segmentation algorithms: ISODATA (Iterative Self-Organizing Data Analysis Technique Algorithm), fuzzy c-means and active contour without edges.

Feature extraction has been made through color and texture descriptors. The most used color descriptors are color histograms and color moments [Wadhawan et al., 2011b; Situ et al., 2008; Barata et al., 2013]. The variety for texture descriptors is vast: wavelet coefficient, Haar Wavelet, Gabor filter, Gray Level Co-Occurrence Matrix (GLCM), Active Shape Model (ASM), for instance. Those descriptors are applied in several works [Wadhawan et al., 2011b,a; Situ et al., 2008; Barata et al., 2013; Doukas et al., 2012]. Most studies reported that better results are reached when they combined color and texture descriptors [Marques et al., 2012].

After feature extraction, many approaches use the low-level descriptors as input for the classifier. On the other hand, some authors improve their methods by processing the low-level information before the classification step. Mid-level feature extraction aims at transforming low-level descriptors into a global and richer image representation of intermediate complexity [Boureau et al., 2010]. The most popular mid-level representation is the BoVW approach. Examples of BoVW techniques on the melanoma classification problem can be found in Wadhawan et al. [2011a]; Situ et al. [2008]; Barata et al. [2013]. They work with color and texture features on the low-level and produce the mid-level by aggregating information via k-means clustering algorithm. The codebook size is small, ranging from 100 to 500 visual words [Barata et al., 2013; Situ et al., 2008]. A more detailed analysis of Wadhawan et al. [2011a] will be given in Chapter 4, since it was chosen as benchmark for our work. This choice

is justified by the use of the same dataset and also because it is one of the most complete works in terms of method description.

Traditionally, classification process have been made with Suport Vector Machines (SVM) [Vapnik, 1995], a very popular and powerful learning technique for data. Many authors applied the SVM classifier [Wadhawan et al., 2011b,a; Situ et al., 2008; Doukas et al., 2012; Mete and Sirakov, 2012; Abbas et al., 2012]. In short, what differs one work from others is the kernel function used on the SVM and its parameters. Also, some authors employed other classification methods, such as neural networks [Iyatomi et al., 2008; Mikos et al., 2012] or decision-trees [Di Leo et al., 2010; Capdehourat et al., 2011]. The experimental validation protocol depends on the dataset size: usually it is done with 10-fold cross-validation, but some studies like Iyatomi et al. [2008]; Mikos et al. [2012]; Marques et al. [2012] use a leave-one-out schema.

Scharcanski and Celebi [2014] also compares the main works cited in this section. This book is a compilation of the last papers published so far about automated melanoma screening. Although the most part of the book is related to dermoscopic image processing, some works deal with clinical images, usually addressing illumination and reflectance problems.

Most part of the literature is focused in border detection, lesion segmentation and meta analysis of the already proposed methods. In the last years, the literature continued to present new melanoma classification works, but always employing the color and/or texture descriptors as image features [Fidalgo Barata et al., 2014; Abuzaghleh et al., 2014; Barata et al., 2014; Abedini et al., 2015]. Barata et al. [2014]; Fidalgo Barata et al. [2014], specially, reinforces the need of lesion segmentation before extracting the image features.

Despite the existence of several works for melanoma classification, they are not directly compared due to the use of distinct datasets and different validation protocols among the methods. Also, there is no official public melanoma dataset to promote different methods experimentation on same conditions.

Table 1 – Results reported in the literature. Table extended from Fornaciali et al. [2014].

Reference.	**Method**	**Dataset #pos/#neg**	**AUC (%)**	**ACC (%)**
[Iyatomi et al., 2008]	Color and texture descriptors; Neural network	198/1060	92.8	*
[Situ et al., 2008]	Color histogram; Gabor filter; BoVW; SVM	30/70	82.2	*
[Wadhawan et al., 2011b]	Color histogram; Haar wavelet; SVM	110/237	*	76.4
[Wadhawan et al., 2011a]	Haar wavelet; SVM	388/912	91.1	*
[Abbas et al., 2012]	ABCD rule-based features; SVM	60/60	88.0	*
[Doukas et al., 2012]	ASM; SVM	800/2200	*	85-90
[Marques et al., 2012]	Color and texture descriptors; *	17/146	*	79.1
[Mikos et al., 2012]	GLCM; Neural network	42/88	*	69.5
[Barata et al., 2013]	Color histogram; Gabor filter; BoVW; k-NN	25/151	**	**
[Abuzaghleh et al., 2014]	Color and texture descriptors; SVM	40/160	*	90.6
[Barata et al., 2014]	Color and texture descriptors; SVM/KNN/AdaBoost	25/151	**	**
[Fidalgo Barata et al., 2014]	Color constancy algorithms; BoVW; SVM	241/241	*	84.3
[Abedini et al., 2015]	Color and texture descriptors; BoVW; SVM	40/160	**	**

AUC: area under the ROC curve | ACC: accuracy | *This information was not reported by the authors in the original paper | **Uses Sensitivity and Specificity, as evaluation measure. The values reported are, respectively: 93%/85% [Barata et al., 2013]; 96%/80% [Barata et al., 2014]; 90%/90% [Abedini et al., 2015]

2.3.2 Melanoma Screening on Handheld Devices

Using smartphones for melanoma screening has several advantages, since those devices are simple to use, the examination does not require complex equipment and acquisition procedures. Besides that, they are also more convenience and low-cost when compared to more conventional computers (e.g., desktop or laptop). Therefore, by offering portability and ubiquitous connectivity, those devices are a powerful help to save lives as noted by Allen [2015].

Wadhawan et al. [2011a] proposed a framework for melanoma screening on handheld devices called SkinScan. The library was made on 2011 but is not available for public use. The authors reported an area under the curve (AUC) of 91%, based on a 10-fold cross-validation on a dataset of 1300 images, being 388 melanomas, that were upload to the phone. They also published other paper [Wadhawan et al., 2011b] implementing the 7-points checklist, and reinforcing the use of color and texture descriptor to extract features. The findings of these authors led to a system improvement able to detect melanoma and other skin lesions using handheld devices. The new results were published in Zouridakis et al. [2015] that can be considered an extension of SkinScan, which is now commercially called SkinVision.

Other similar studies [Doukas et al., 2012; Mikos et al., 2012] analysing skin lesions on smartphones. Both have similar main user case: the user photographs the abnormal skin and annotates the lesion. The application extracts features, analyses them and returns a diagnoses if it is a melanoma or not. Doukas et al. use the WEKA SVM to classify the features, as Mikos et al. prefer neural networks. Another difference between these works is the dataset size: Mikos has 130 images, being 42 melanomas while Doukas made the experiments on 3,000 images being 800 melanomas. This difference reflects on the results, since Doukas achieved an accuracy of about 87% while Mikos had just 70%. To finish this comparison, another contribution of Doukas is the use of cloud computing to process the images, allowing the pipeline to be used for different Operating Systems, like Android, iOS and Windows Phone regarding the differences between them.

There are other products for automated melanoma screening directed to community, patient and generalist clinician users. A good review of these applications can be found in Kassianos et al. [2015]. Although these new technologies offer the promise of improving early melanoma detection, they often present low accuracy on their results [ISI, 2015; Wolf et al., 2013].

Recently, the survey proposed by March et al. [2015] reassesses the past works, introduces other commercial applications and also discuss the regulation of mobile technologies for medical purposes. Nevertheless, this study also concluded that, at this time, there is no mobile system completely accurate to be used in melanoma screening. However, the authors

highlight that no experiment was done in a true clinical setting in order to compare the performance of dermatologists using or not any of these new automated techniques.

2.3.3 Quality Analysis of Current Methods

When we talk about the quality of current works in the literature of melanoma screening, there are two points of view: the physicians' and the computer scientists'. Physicians are usually worried in analysing new systems as complete tools to be used in screening programs as a decision support mechanism, so they search for sensitive methods with high accuracy. Those, however, are not commonly found in academic literature, since most of works are still small studies, works in progress. When the system is implemented on mobile devices and reach final users directly, physicians are, rightfully, very critical about the quality of the methods, since inaccurate systems can mislead patients to a false sense of security with a false negative outcomes [Tyagi et al., 2012; Wolf et al., 2013].

On the other hand, computer scientists understand that systems must have high accuracy rates to be used in screening programs, but they are also excited about the improvements of their methods along time, which justify publications with AUC between 80–90%. Nevertheless, the works in literature are not directly comparable among themselves, since they employ different datasets that are not public, and are very hard to obtain even under request.

Another important matter about the quality of current works is the question of reproducibility. Since works are not directly comparable, new researchers need to reimplement previous literature from scratch in order to compare new approaches to existing ones. As if this inversion on the "onus of reproducibility" were not bad enough, existing methods and protocols are often described so cursorily as to prevent any attempt at all of reproducibility.

2.4 Conclusion

In this chapter we introduced the automated melanoma screening problem by the optics of Computer Vision. For this, we visited the main techniques of image classification found in the literature, addressing its characteristics, advantages and disadvantages. Our attention was driven to models based on BoVW, since they have some advantages that are suitable for the melanoma screening problem: (i) they don't require huge amounts of images to construct the predictive model, (ii) the approaches can be extended by other techniques, improving the results, and (iii) there are several works in literature that can benefit from the use or modernization of this technique.

Our literature review covers the main works of automated melanoma screening by im-

ages, pointing out their similarities, differences and main results. We also described melanoma screening in handheld devices, reinforcing the importance and the potential for exploitation of this issue.

Chapter 3 describes our solution. It is based in the cutting-edge representation BossaNova [Avila et al., 2013], one of the most recent extension of the BoVW model. Its conceptual and practical details will also be explored in the following chapter.

3 Proposed Solution

In Chapter 2, we have introduced the state-of-the-art on automated melanoma screening. Under the point-of-view of Computer Vision, it is an image classification problem, so our literature review also included the most relevant works about this theme. We have seen that the BoVW model is one of the most successful models for image classification, especially when the amount of annotated images is scarce, which is the case for medical images. This model has been improving in the last twelve years and now we can find advanced approaches that improve the classification rates by the cost of generating huge descriptors that consume more computational resources.

The melanoma classification literature has a typical protocol composed by three steps: (i) lesion segmentation, (ii) feature extraction inside the lesion and (iii) classification itself. What differentiates one work to others is the feature descriptors used in the experiments: while one author will opt to use certain texture or color descriptors, another will prefer a different choice. The classifier is generally chosen among a small handful of choices, SVM being very common. Nevertheless, these techniques are often reductive, since they explore poor feature descriptors, simple schemes of coding and/or poling, small sizes of the codebook and a mid-level representation that do not incorporate relevant aspects of the visual content.

In such a way, a critical review of the current literature is lacking. In this sense, this work opens the opportunity to advance the state-of-the-art by probing cutting-edge BoVW extensions. Section 3.1 presents our solution for automated melanoma screening and Section 3.2 describes our spatial pooling strategy specially designed for this problem. Besides that, there are open problems yet not explored in the literature, which are described in Section 3.3. Finishing this chapter, Section 3.4 summarizes the information presented here.

3.1 A Modern Approach for Melanoma Screening

A preliminary analysis of related work of melanoma classification indicates that the most serious problem of literature is the use of simplistic techniques, like outdated BoVW models or worse, which possibly do not exploit the full potential of the images in the composition of low and mid-level descriptors. It is known, however, that the model was enhanced and today achieves good results on diversified image classification tasks. Thus, the main contribution proposed by this work is the use of modern BoVW based techniques in the automatic screening of melanoma. It is clear that enhanced mid-level descriptors were not explored yet,

opening the opportunity for further investigations and improvements.

Among the advanced approaches of the BoVW model, we opt to employ BossaNova as the basis of our framework since it has been showing competitive results that overcome the state-of-the-art for image classification tasks. The original contribution of this work is a novel application for BossaNova: this is the first time that it is applied to melanoma classification.

BossaNova is an enhanced mid-level representation that brings several novelties for the melanoma screening problem using BoVW model. Proposed by Avila et al. [2013], the contributions are present in both low- and mid-level stages. Among the low-level advantages, we highlight robust descriptors with dense schemes for sampling and reductions of dimensionality to speed up the process. The innovations in the mid-level feature extraction are new sizes for the codebook, a soft coding schema, a density function-based pooling strategy, normalizations of the final feature vector and also incorporation of spatial information. All of these contributions will be detailed next. The main pipeline of melanoma classification using BossaNova is shown in Figure 12.

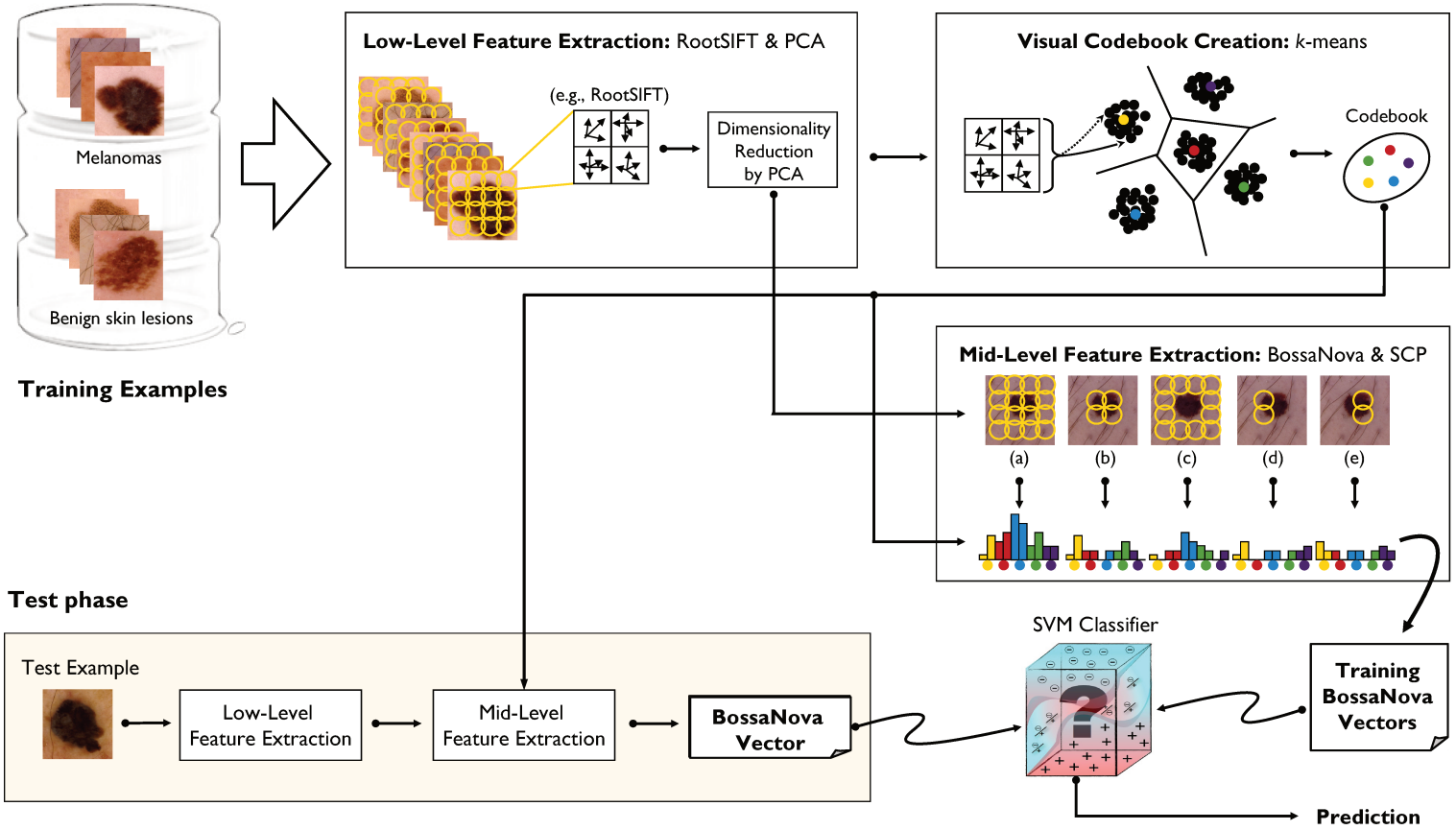

Figure 12 – The main pipeline of BossaNova. **Low-Level Feature Extraction:** RootSIFT descriptors [Arandjelovic and Zisserman, 2012], which yields superior performance than SIFT, are employed. The dimensionality of the RootSIFT is reduced from 128 to 64 by using PCA. The PCA matrix is learned over a sample of low-level features during the training phase. **Visual Codebook Learning:** During the training phase, k-means with Euclidean distance is run over a sample of one million low-level features, the final centroids are used as codewords. **Mid-Level Feature Extraction:** BossaNova descriptors creates the feature vectors for the images. The spatial pooling takes into account tone of the pyramid schemes, creating one independent feature vector for each hierarchical region of the pyramid and then concatenating them. **Decision Model Training:** During the training-phase, the BossaNova vectors of annotated images are employed to train a decision model using SVM. **Decision Model Prediction:** The trained model employs the BossaNova feature vectors of an image to predict on the positive (melanoma) or negative classes.

Following the common pipeline of image classification using BoVW model, the first stage is the **low-level feature extraction**. Our approach has advantages in three aspects:

- **Robust descriptors**: while other authors use simple features to describe the images (like color and texture descriptors), here we propose the adoption of the robust descriptor RootSIFT which yields superior performance without increasing processing or storage requirements that SIFT already needs [Arandjelovic and Zisserman, 2012].
- **Dense sampling**: some authors decide to extract features only related to points-of-interest of the images (like corners or edges, for example). Here we employ RootSIFT in a dense sampling strategy, that is, describing the whole image in order to extract as much information as we can. Figure 13 compares the two types of low-level extraction.

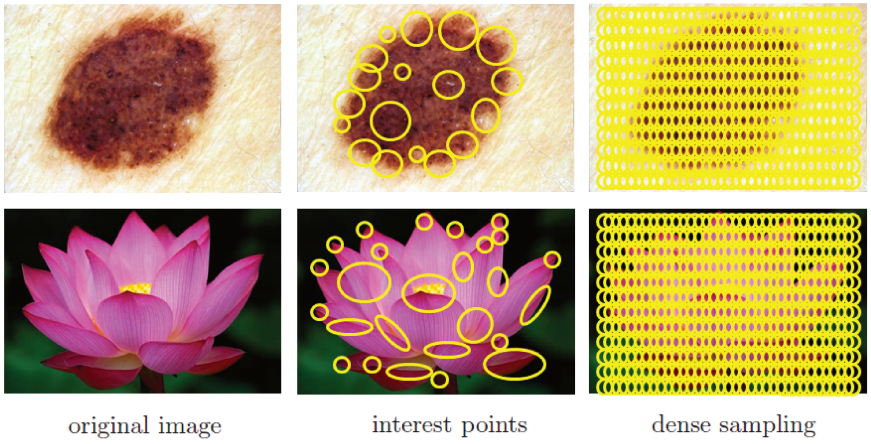

Figure 13 – Types of low-level extraction. First with points-of-interest and then with dense sampling. Figure adapted from [Tuytelaars, 2010].

- **Reduction of dimensionality**: finally, since the RootSIFT are extracted in a dense sampling way, the final low-level features are usually bigger than the images themselves in terms of storage. So, in order to accelerate the processing of this huge amount of data, we apply PCA to reduce the dimensionality of the low-level descriptors.

Moving to the **mid-level feature extraction**, the main contribution of our approach is the use of BossaNova as enhanced descriptor. Besides that, since we are generating mid-level features of up to thousands dimensions, it is convenient to explore new sizes of codebook, what will be shown in Chapter 4. BossaNova brings improvements in four aspects:

- **Coding**: the classical BoVW model employes hard assignment for coding, that is, it assigns each feature vector to the closest codeword of the codebook. BossaNova uses a

soft-assignment strategy. It was chosen due to better results without prohibitive computational costs [Yang et al., 2009; Boureau et al., 2010]. The soft-assignment associates each feature vector to the K-nearest codewords and the values of the attributions are related to the Euclidean distance of the feature to the codeword subjected to the standard deviation of the cluster in question.

- **Pooling**: instead of using the classical sum- or max-pooling strategies, BossaNova introduces a density function-based pooling schema, aggregating local spatial information about the descriptors around each codeword, preserving thus statistical information about the distribution of the features. It is done by computing a local histogram z of distances between the descriptors found in the image and those in the codebook. The intuition is shown in Figure 14.

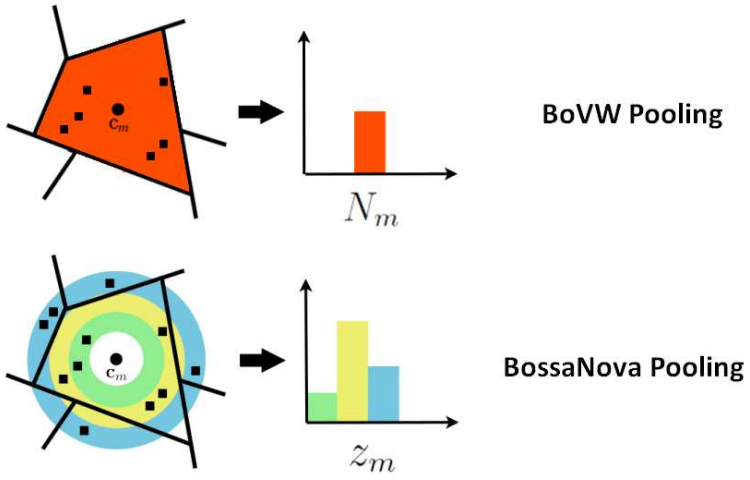

Figure 14 – BossaNova's intuition. Reproduced from [Avila et al., 2013].

While in the standard BoVW pooling all descriptors close to a codeword are quantized by the same histogram bin, in BossaNova the descriptors are quantized in different histogram bins accordingly to its distance around the codeword. The main advantage of this new scheme is that it preserves the information about the descriptors' distribution around each codeword. Another advantage is that the degree of information preserved can be adjusted by the number of bins of the z histogram. From Figure 14 it is easy to see that when the number of histogram bins in BossaNova is equal to 1, we have the BoVW pooling approach. This illustrates that BossaNova is, in fact, an extension of the BoVW model.

To construct the local histograms, BossaNova uses the parameters shown in Figure 15. BossaNova vector is defined by three parameters: the number of codewords M, the

number of bins B in each histogram, and the range of distances α_{MIN} and α_{MAX}. The former, α_{MAX}, avoids considering words not close enough from the center and α_{MIN} avoids the empty regions that appear around each codeword, saving space in the final descriptor. The BossaNova Z is a vector of size $M \times (B + 1)$;

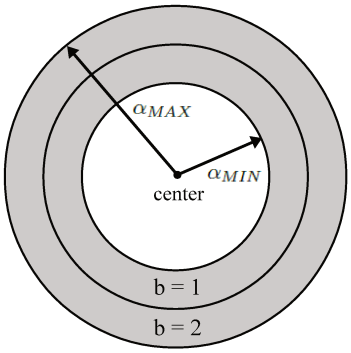

Figure 15 – Illustration of the range of distances α_{MIN} and α_{MAX} parameters of the BossaNova model. α_{MAX} avoids considering words not close enough from the center and α_{MIN} avoids the empty regions that appear around each codeword, saving space in the final descriptor. The gray area corresponds to the bounds of the histogram, local descriptors outside those bounds are ignored. Figure adapted from [Avila et al., 2013].

- **Normalizations**: normalization is the adjustment of values measured on different scales to a common range. When the number of codewords increases, the local histogram becomes sparser. Besides that, Perronnin et al. [2010] observed that when the feature vectors become too sparse, the similarities become less reliable. To attenuate this phenomenon, they proposed a power-law normalization. BossaNova does the same by taking the square root of each histogram bin. Besides that, BossaNova applies a ℓ_2-normalization to rescale the final vector.

- **Spatial information**: Lazebnik et al. [2006] proved that the incorporation of spatial information of the features along the images can improve BoVW models. Their SPM is one of the most used schemes in existing classification methods. Here, BossaNova also employs this schema grouping feature vectors according to their location in the image (for example, top left corner, bottom right corner, and so on).

For all the improvements mentioned above, we believe that our approach will deliver better results for melanoma screening than the literature. We believe, still, that our solution based in the cutting-edge BossaNova mid-level representation will be robust to noise in the images, dispensing any ad-hoc image pre-processing technique like lesion segmentation and hair removal illustrated in Figure 16.

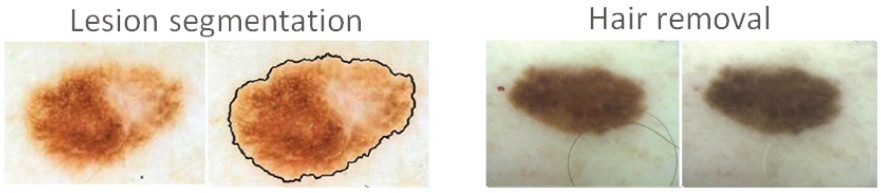

Figure 16 – The most common image pre-processing techniques applied in melanoma screening works. The lesion segmentation and hair removal are usually done ad-hoc of the proposed methods before feature extraction. Figure adapted from Di Leo et al. [2010]; Capdehourat et al. [2011].

3.2 Spatial Circular Pooling

The ABCDE rule [Nachbar et al., 1994] and the 7-Points Checklist [Argenziano et al., 1998] are pioneers methods for melanoma screening through dermoscopy in the physicians' community. They evolved to modern approaches like the 3-Points Checklist [Soyer et al., 2004] and the 7-Points Checklist Revisited [Argenziano et al., 2011]. Since the original methods were extensively tested and there is a consensus about their effectiveness, many researchers of automated melanoma screening, based upon Computer Vision techniques, try to reproduce these steps in computers. These rules are related to appearance and morphological aspects that are simulated with color and texture descriptors. This kind of image descriptors has been used as the basis for existing melanoma classification systems.

Our method, on the other hand, uses robust descriptors that do not require color and texture details of the images, but can be enriched by information about the spatial distribution of features. However, some aspects of these rules, like asymmetry and border shape, are also related to spatial information, so we believe that such medical rules can also improve the classifier in advanced BoVW based approaches. Furthermore, not all melanomas follow the ABCD rule or any other classification pattern, so it is important to validate if these methods can be more informative with the computer vision techniques that we have today.

In the literature, the authors capture spatial information about the skin lesion by segmenting it and analyzing the border, before extracting the image features. This procedure can be time consuming and prone to errors, what motivates new measures to evaluate border detection [Celebi et al., 2009]. This leads us to two questions: (a) is the segmentation really important for melanoma screening? (b) if we separate the image features of the lesion and healthy skin would we have a better classifier?

To investigate it, we started using the PH2 Dataset[1] [Mendonca et al., 2013] since

[1] **PH2 Database**: created by Mendonca et al. [2013], this dataset contains a total of 200 dermoscopic

it has information about the lesion segmentation. We performed the following experiment divided in two approaches: (1) extract features from the whole image, (2) extract features only inside the lesion. For both approaches, we used BossaNova to construct the mid-level representation and compare them in order to identify which one leads to a better melanoma classifier. Using a 5-fold cross-validation schema, we have found that approach (1) leads to a classifier with an accuracy of 87%, while approach (2) leads to a classifier with an accuracy of 88.5%. This motivated us to further explore the spatial information of the lesion without need to segment it, leading to improved results without incurring the computational costs and possible errors of automated segmentation.

Therefore, we implemented a brand new pooling strategy named SCP. SCP is a type of spatial pooling addressed specially for the skin lesion classification problem. It enriches the BossaNova representation by adding spatial information about the image descriptors distribution around the skin lesion. Also, the SCP can be extended to other BoVW based techniques in a straightforward manner.

SCP is a new, fast and easy way to extract the lesion without need to segment the image. It was designed observing that, typically, dermoscopic images are concentric, that is, the lesion is centered on the image and it occupies about 50% of the image area. The method is explained in Figure 17 (top row): we draw a circular region with radius R to capture 50% of the image area (see Equation 3.1), we consider 5 sampling vectors composed by (a) the whole image, (b) the outer and (c) the inner regions and (d) the left and (e) the right sides of the lesion. The schemes (a)-(c) try to evaluate the impact of lesion segmentation over the classification, and the schemes (d)-(e) try to identify asymmetrical borders, which is a relevant criteria according to the ABCD rule of dermoscopy [Nachbar et al., 1994].

$$R = L/\sqrt{2\pi}, \tag{3.1}$$

where R is the radius of the circle used on the SCP approach and L is the size of the square skin lesion image. From Figure 17 is easy to see that while SCP tends to emphasize the contrast between the center and the border of the image, SPM tends to emphasize the contrast between its quadrants. While we expected the center–border contrast to be very important (because it corresponds to the rules of dermoscopy image analysis), our results show that SPM and SCP perform equally.

SCP is, therefore, an attempt to incorporate medical rules for dermoscopic melanoma

images with 768×560 pixels of resolution, being 80 common nevi, 80 atypical nevi, and 40 melanomas. The dataset includes annotations and segmentation of each lesion, that were all obtained at the Dermatology Service of Hospital Pedro Hispano (Matosinhos, Portugal). This dataset can be found at: http://www.fc.up.pt/addi/ph2%20database.html.

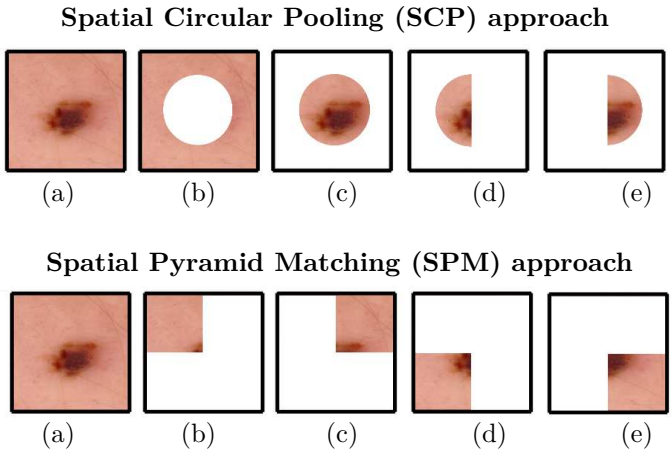

Figure 17 – Comparison between the SCP (top row) and the SPM approaches [Lazebnik et al., 2006] (bottom row), contrasted in our evaluation of the factors affecting the model accuracy.

classification (like ABCDE rule and 7-Points Checklist) in advanced BoVW-based systems for automated melanoma screening.

3.3 Other Questions Investigated by This Work

As mentioned in the beginning of this chapter, we have other questions to evaluate besides how modern approaches for the BoVW model perform in the melanoma screening problem.

Since the BoVW model is divided in three main steps (low- and mid-level feature extraction and classification), it is important to determine how much each of the feature extraction steps influences the accuracy of the classifier. This is particularly import to guide future researches in this field.

Other relevant aspect to be exploited is how the size and the quality of the training set impact the accuracy rates. It is important to investigate it because since each author reports his/her results in different datasets, maybe the methods found in literature are not so different in terms of performance, but the differences came from bigger amounts of samples being imputed into the classifier.

Finally, this work presents a set of methodological contributions that can benefit the melanoma screening community in order to make researches reproducible and easily comparable. These questions will be fully explained in Chapter 4 - Experimental Results.

3.4 Conclusion

This chapter introduced our solution for automated melanoma screening. It was seen that the literature counts with other BoVW-based approaches for this problem, but they are not applying modern improvements of such model.

This opens the opportunity to investigate how enhanced mid-level features perform in this particular case of medical imaging. It is also the first use of BossaNova descriptors in this context.

In addition, we introduce, the SCP as a special schema to incorporate spatial information of the image features, in a fashion inspired the ABCD Rule, but without the rigidity imposed by segmentation. As it was presented, there are several justifications about the problem and medical knowledgement that led to believe that the SCP would have a positive impact in the accuracy of the classifier but, as will be seen in the next chapter, the experimental data didn't confirm this hypothesis. Chapter 4 details the experimental design and also analyzes the possible causes of this phenomenon.

4 Experimental Results

This chapter shows our empirical results. It is organized as follows: Section 4.1 introduces the datasets used in the experiments. Section 4.2 describes the evaluation metric, justifying our choice. In sequence, the set of experiments are detailed in Section 4.3 in terms of its objectives, experimental design, results and analysis. By its time, Section 4.4 concludes this chapter and summarizes the main findings.

4.1 Datasets

All experiments described in this study are related to one of the datasets listed below.

1. **IRMA Dataset**[1]: created by the Department of Medical Informatics of the RWTH Aachen University, this dataset is composed of 747 dermoscopic images of skin lesion with resolution of 512×512 pixels, being 187 melanomas and 560 benign skin lesions.

2. **Interactive Atlas of Dermoscopy**[2]: created by several researchers from Italy and Austria, the interactive Atlas of Dermoscopy is a multimedia project for medical education. It contains a CD-ROM with over 2,000 images of pigmented skin lesions, divided into dermoscopic or clinical ones, including its diagnosis and histopathologic data.

From the Computer Vision point-of-view, the main challenge of this research are the similarities between the classes of the images being classified. Although they were extracted from IRMA Dataset, Figure 18 represents the dermoscopic images of any melanoma dataset. Note that melanomas (top row) and benign skin lesions (bottom row) are very similar. Other challenges are smooth transitions between the lesion and normal skin, making difficult lesion segmentation, and the occlusions caused by hair.

4.2 Evaluation Metrics

For all experiments, we used the Area Under the Curve (AUC) as evaluation metric. The AUC is the area under the Receiver Operating Characteristic (ROC) Curve. This curve is a

[1] IRMA Datasets - http://ganymed.imib.rwth-aachen.de/irma/datasets
[2] Interactive Atlas of Dermoscopy - http://www.dermoscopy.org/

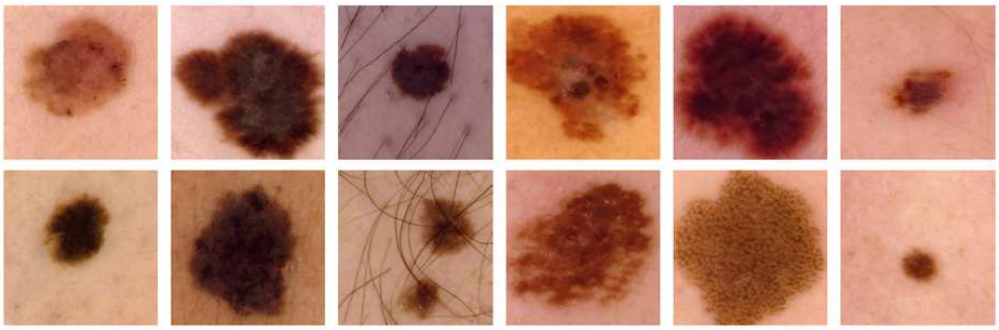

Figure 18 – Melanoma images (top row) and benign skin lesions (bottom row).

graphical representation of the performance of a binary classifier, tracing the True Positive Rate (Senstivity) in function of the False Positive Rate (1−Sensitivity).

We chose the AUC as evaluation metric because it provides a global measure of the method being evaluated, without taking into account the precise choice of operating points, i.e., the exact compromise between sensitivity and specificity preferred by the user. While sensitivity and specificity can be balanced according to the cost of the problem, the AUC gives a more global evaluation of the method. This also explains its popularity in the literature of automated melanoma screening.

4.3 Experiments

Before performing the experiments of melanoma screening described in the following, we started with a simple experiment trying to reproduce the same results reported by the authors of BossaNova for the Oxford Flowers-17 dataset[3]. This experiment aims to check if we are using BossaNova framework in the right way. We achieved the same results in this dataset that the authors, so it suggested that our pipeline was correct. The Flowers dataset was chosen because it shares *some* characteristics with the melanoma screening: it's an special purpose dataset, trying to differentiate between similar classes.

For all experiments the classification was performed by Support Vector Machines . We always used the popular LIBSVM library [Chang and Lin, 2011].

4.3.1 BoVW × BossaNova

One of our main contributions is the use of an enhanced BoVW-based technique in the melanoma screening problem. Advances in the last five years indicate that recently exten-

[3] Oxford Flowers-17: http://www.robots.ox.ac.uk/ṽgg/data/flowers/17/

sions of the BoVW model are highly supposed to deliver better results in any classification problem. So we are also interested in investigating how these extensions perform for melanoma classification. The first experiment of this research is, therefore, a comparison between traditional BoVW and BossaNova[4] in the same classification task. We opt to use BossaNova as the starting point of our framework due to its performance, comparing well with the state-of-the-art for several challenging datasets of image classification [Avila et al., 2012, 2013].

Goal: to investigate if the elected advanced BoVW-based framework for general image classification is also suitable for specific datasets, especially for melanoma ones.

Experimental design: we used the IRMA Dataset in a 5-fold cross-validation schema in order to eliminate deviations introduced by random choices of the images for training or testing. To be fair, we adopted the default parameters of each method, without tuning them. When a parameter is common for both approaches, we used the same value in both experiments. The parameters are: codebook size ($M = 1024$), alphas ($\alpha_{MIN} = 0.4$ and $\alpha_{MAX} = 2.0$), number of bins ($B = 2$), number of neighbors ($K = 10$).

Results: as expected, that BossaNova leads to better results than traditional BoVW approaches in the melanoma screening problem. While BoVW reported an AUC of 89.45%, BossaNova achieved 91.51%. A Student t-test [Jain, 1991] indicates that BossaNova is better than traditional BoVW with a confidence of 90%.

Analysis: the experiment proved that enhanced mid-level features achieved better results than the traditional BoVW model. Although the difference seems to be small, in practice BossaNova outperforms BoVW with over 2% of absolute improvement.

Once the experiments were performed without sophisticated parametrizations, we can accept the hypothesis that modern extensions of the BoVW model are more adequate for melanoma screening.

4.3.2 Low-level × Mid-level: Which Influences More the Classification Performance?

Image classification with BoVW-based models can be decomposed in two sequential steps: the low-level and the mid-level feature extractions. Each BoVW approach has its singularities, especially for the parametrization of theses steps. So, a good clue to improve the classification rates is to investigate which combination of the parameters values leads to better results. Since

[4] In our experiments, we used the BossaNova code available at https://sites.google.com/site/bossanovasite/.

the number of combinations can be exponential and the experiments are usually slow, this kind of analysis is not common in literature. On the other hand, a hint on which step should be better exploited is highly desired. This is still lacking and could guide future researches in this problem.

We studied the impact of the low-level and the mid-level over the final classifier. The study was performed via a *fractional factorial design*[5] in order to evaluate both the significance and the relative importance of each evaluated factor in the results. No previous BoVW-based work in the literature performed an evaluation as broad in scope as ours, neither as rigorous in terms of statistical design.

Goal: to identify which feature extraction step is more important and significant for the classifier accuracy. Besides that, identify which BossaNova's parameters are more relevant for melanoma screening problem.

Experimental design: we have chosen six attributes of the framework, analyzing their contribution for the final result. The low-level attributes are step and size values of the RootSIFT extraction [Arandjelovic and Zisserman, 2012]. The mid-level attributes are the codebook size, the number of bins, the maximum and minimum values for $\boldsymbol{\alpha}$ parameter and the pooling schema (see next experiment). Each combination was validated in a 5-fold cross validation schema using the IRMA Dataset, whose images were re-sized to 316 $\times$ 316 pixels due to minimize time and computational resources consuming. The setup values are detailed on Table 2. The Analysis of Variance (ANOVA)[6] was employed to analyze the differences between the averages of each group.

Results: the fractional factorial statistical results are shown on Table 3. We omitted the second-order interactions since none of them were significant. On the other hand, all main effects were significant. The choices of step and scale for the low-level and the choice of number of bins and codebook size for the mid-level explain most of the non-random variation, as seen in the Sum of Squares column. Besides that, the residuals contain most of the information about variability in the classification. This indicates that no parameter combination has systematic large advantage throughout all 5 folds.

[5] A *full factorial experiment* is an experiment with two or more parameters, in which each one has a finite set of values to be evaluated. All parameters' combination are exploited leading to an exponential number of validations. A *fractional factorial design* is an experimental setup on which a subset of experimental runs of a full factorial design are carefully chosen, trying to expose information about the most important features of the problem (see [Jain, 1991, chap. 17] for more details).

[6] *Analysis of Variance* (ANOVA) is a statistical model used to analyze group averages and their variations in order to identify significative differences between them and if the parameters influence any dependent variable (see [Jain, 1991, chap. 15] for more details).

Therefore, the random fluctuation across the folds appears to be very important, suggesting that the choice of the training set affects very much the results. In order to stabilize this fluctuation, each fold should be reasonably balanced, not only on the number melanoma and non-melanoma images, but also in other factors like image quality (hair presence or not, good and bad illumination, and lesion size) and types of lesions in the negative class. Despite the large residuals, the main effects that came significative were for step and min scale (p-value < 0.001), as well as the number of bins and the codebook size. We conclude, then, that these parameters are good clues to be explored when constructing a BoVW-based method for melanoma classification.

Analysis: the statistical analysis still shows that the low-level has bigger impact over the classification than the mid-level. It can be proved by the column 'Sum of squares': note that the low-level concentrates higher values for this parameter. This shows that in order to construct a good melanoma classifier, we should pay attention on the feature extraction step. Although less informative than the low-level, the mid-level also plays an important role on the classification. Remarking that the effect of the codebook size was very significative (p-value $= 6.6 \times 10^{-4}$), this shows that the choice of the codebook size significantly improves the predictive power of the model. In addition, an analysis of the ANOVA table shows that the step choice was, arguably, the most influential factor (largest partition of the mean square variation), reinforcing the relevance of that factor on improving the classification model.

This experiment revealed that the low-level feature extraction is the most relevant step for image classification with BoVW-based methods. This is expected since the low-level features feed the subsequent steps of the pipeline. This result indicates that researchers should pay attention to the image descriptors employed in their solutions. We also highlight that the BossaNova parametrization can impact the classification rates as pointed by its authors. In the particular case of melanoma screening, the codebook size and the number of bins in the histogram are the most important parameters. These findings are essential for driving future investigations.

4.3.3 Spatial Circular Pooling

Our second main contribution is the proposal of a new spatial pooling strategy especially designed for melanoma screening. The intuition comes from previous works in image classification literature that indicates that the accuracy of BoVW models tend to improve just by incorporating spatial information of the images into the bags. The main traditional spatial pooling operation is the pyramids proposed by Lazebnik et al. [2006]. It works very well for

Table 2 – Parameters of the Fractional Factorial experiment. Each line shows a parameter of the experimental setup, tested with high and low values, in order to identify its impacts on the classification result. Table reproduced from Fornaciali et al. [2014].

Parameter			Value
Low-level	**step**	*small*	8
		big	24
	scale min	*min*	12
		max	24
	scale max	*min*	64
		max	128
	#scale	*few*	2
		many	4
Mid-level	**#bins**	*few*	2
		many	4
	α	*tight*	[0.6, 1.6]
		loose	[0.2, 2.0]
	codebook	*small*	1024
		big	2048
	pooling	SPM [Lazebnik et al., 2006]	1×1+2×2
		SCP (ours)	(see Fig.17)

Table 3 – Partial view of the ANOVA Table. Table reproduced from Fornaciali et al. [2014].

Level	Parameter	Degrees of freedom	Sum of squares	Mean square	F value	p-value	
Low-level	**step**	1	40.66	40.66	110.76	$< 2.00 \times 10^{-16}$	***
	scale_min	1	9.64	9.64	26.26	5.45×10^{-7}	***
	scale_max	1	1.60	1.60	4.35	3.79×10^{-2}	*
Mid-level	**#bins**	1	4.85	4.85	13.21	3.29×10^{-4}	***
	codebook	1	4.35	4.35	11.85	6.60×10^{-4}	***
-	**Residuals**	291	106.84	0.37	-	-	

Significance codes: *** p-value < 0.001; ** p-value < 0.01; * p-value < 0.05

general purpose classification. Our motivation was to develop a spatial pooling operation that considers important aspects of medical knowledge, for example, by aggregating information inspired in the ABCD rule. Other authors of melanoma screening techniques tend to reproduce the ABCD rule just by using color and/or textual descriptors for the low-level feature extraction. Since we adopted advanced descriptors (like RootSIFT) that don't deal with color and/or textual aspects of the images, the ABCD characteristics were incorporated in other way.

Goal: to investigate if the SCP is more suitable for melanoma screening than other spatial

pooling strategies.

Experimental design: this experiment is a subset of the "Low-level × Mid-level" examination. The key parameter here is the spatial pooling schema. SCP was validated against SPM proposed by Lazebnik et al. [2006]. For a graphical comparison between these two types of spatial pooling, see Figure 17. Since the images were re-sized to 316 × 316 pixels each one, the radius R of Equation 3.1 is 126.1 pixels, starting at the center of the image.

Results: the results of this experiment are also related to the ANOVA shown in Table 3. Contrary to what we expected, the choice of the spatial pooling strategy was not significant enough to cause major impacts on the classifier, which justifies why this parameter is not listed in the ANOVA Table. Our analysis showed that the average AUC for both SPM and SCP approaches is 93.7%, and the ANOVA did not show statistical differences between each approach. In order to better visualize the similarities of both approaches, Figure 19 presents the ROC curves for SPM and SCP.

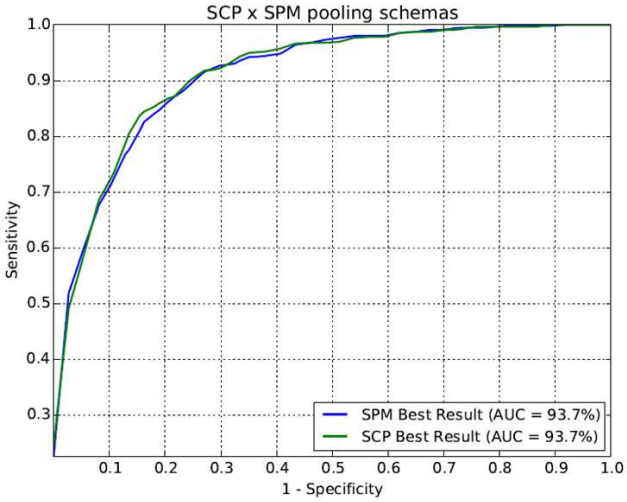

Figure 19 – Best ROC curves for SCP and SPM pooling approaches.

Analysis: the fractional factorial experiment and ANOVA demonstrate that the spatial pooling strategies are equivalent. This suggests that, for melanoma classification problem, both approaches offer the same results and the proposed SCP schema is as good as that one introduced by Lazebnik et al. [2006] (SPM). It is important to note, and it is possible to see in Figure 17, bottom row, that the SPM pooling schema captures information about the asymmetry of the lesion by grouping the descriptors in four regions. The SCP schema, on the other hand, just captures information about the borders and

the center of the lesion, clustering descriptors in two groups: one with descriptor belonging to the lesion and other with descriptors that do not belong to the lesion. We can also conclude that the segmentation proposed in Figure 17, top row, frames (d) and (e) are not sufficient to capture the whole asymmetry of the lesion on the SCP approach. These evidences suggest that in the melanoma classification problem the investigation of the lesion asymmetry is as important as its segmentation.

In order to compare our result with the state-of-the-art, we have constructed the Table 1 that resumes the information presented in Chapter 2. It also shows important aspects about each study, like the dataset size, the proportion between positive (melanoma) and negative (non-melanoma) images and the evaluation criteria: the AUC or the accuracy. We will compare our results only with studies that use AUC as evaluation measure, since we consider it more informative than the accuracy. Our method presents an AUC up to 93.7%. This is directly better than Wadhawan et al. [2011a]; Situ et al. [2008]; Iyatomi et al. [2008]; Abbas et al. [2012]. Also, it should be mentioned that Iyatomi et al. [2008] has border detection and feature selection, forcing a non-natural improvement of the method. In our experiments, we used the whole dataset, without removing difficult cases for a machine classifier, like images with poor quality, excessive presence of hair or if the lesion is not whole fitted on the image. We also do not detect lesion borders, remove hair, improve the contrast between melanoma and non-melanoma skin nor do any other ad-hoc pre-processing to benefit the classifier.

Concluding this section, this experiment was designed in order to evaluate our new spatial pooling strategy (SCP) compared it with the most popular spatial pooling approach (SPM). The images were divided on regular grids creating a pyramid of pooled features with the same dimensionality for both SPM and SCP, that is, the feature vectors will both have the same size (five regions). This comparison aims to identify the informative power of each spatial pooling strategy, i.e., given a feature vector of the same size, we aimed to detect which approach preserves more information about the image. It is straightforward to note that the more information the feature vector has, the better is the classification. Our results showed that there is no statistical difference between the spatial pooling schemes and the SCP, as was designed, was not able to capture important aspects of the lesions. This is a good clue to be investigated in further examinations.

4.3.4 The Impact of the Training Set Size Over the Classifier

Early studies of the literature report experiments being done in different setups in terms of number of images used in the training set and proportion between melanoma and non-

melanoma images in the dataset. This leads us to a question: what is the impact of the training set size over the quality of the classifier? To answer this question, we designed the experiment detailed below.

Goal: to verify if the increase of the training set size improves the classifier.

Experimental design: we used the IRMA dataset separating 20% of the images to compose the testing set. The remaining images were organized in five experiments using different training set sizes (10%, 20%, 30%, 50% and 80%), always including the previous sets to compose the next one. The graphical representation of the experimental design is illustrated in Figure 20. This setup was repeated three times in order to attenuate the influence of the random distribution of the images along the experiments. The BossaNova parametrization was the one that led to the best AUC for the SCP: codebook size ($M = 1024$), number of bins ($B = 4$), and alphas values limited between $\alpha_{MIN} = 0.2$ and $\alpha_{MAX} = 2.0$.

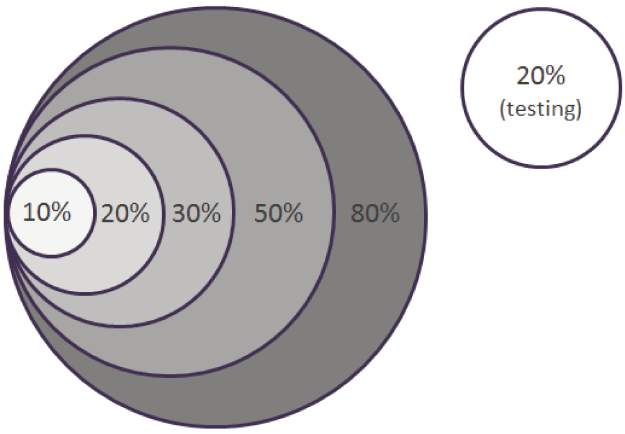

Figure 20 – Experimental setup to evaluate the impact of the training set size over the classifier.

Results: the results are shown in Table 4. Its interpretation is straightforward and will be analised in the next topic.

Analysis: Table 4 shows that, as expected, the random choices of the images for the folds are responsible for fluctuations in the AUC values. This is proved by the the analysis of a same line along the columns. However, this experiment must be interpreted by the average of the three runs. We can see that the average is improving as the training set size increases.

Table 4 – Results for the impact of the training set size.

Size (%)	1st Run (%)	2nd Run (%)	3th Run (%)	Average (%)
10	92.19	86.92	88.08	89.06
20	92.01	88.28	90.04	90.11
30	92.72	87.85	90.62	90.40
50	93.99	89.93	91.26	91.73
80	93.84	91.39	90.92	92.05

This experiment validates our hypothesis that when the classifier is fed with more samples, the classification rates get better. Thus, new approaches to melanoma screening should be validated whenever possible with bulky datasets. It is interesting to note however, that the improvement tends to saturate after some point, suggesting that the sensitiveness to training set size is not as big as for other models (like Deep Neural Networks).

4.3.5 Robustness Analysis

Previous experiments indicated that our method is able to generate a melanoma image classifier with an AUC of 93.7%. The last experiment also revealed that the training set size is a significant factor in the evaluation of methods. In particular, studies that use small datasets (less than 200 images) are subject to overfit, leading to misinterpretation of the learning power of such approaches. The 'Low-Level × Mid-Level' experiment pointed out that despite the importance of the use of enhanced mid-level descriptors, the image variability among the folds is still what most impacts the classifier. So balance the folds in terms of melanoma and non-melanoma images and morphological aspects such as lesion size, lesion centralization, the presence of hair, among others is of paramount importance for the evaluation of automatic melanoma classification systems.

Nevertheless, the last experiments were performed using the IRMA dataset that only contains information about the final diagnosis (melanoma or not). In order to provide a fair balance of images among the folds, deeper details of the images are required. So, the evaluations done in this section employed the Atlas dataset since it provides clinical details of each image, like the type of skin lesion (basal cell carcinoma, blue nevus, melanoma, and others) and the difficulty for a physician correctly classify it. This dataset is also bigger than IRMA and was used in other works of melanoma screening, allowing a less subjective analysis of the proposed method against the literature.

Goal: the main goal of this section is to investigate the robustness of the proposed method in relation to the disturbances that images may present and their degree of difficulty in

a medical analysis. For this, the investigation will be divided into three sub-experiments always doing the best effort to guarantee that the folds are balanced for lesion type and difficulty of classification:

1. Using the complete dataset removing difficult lesions;
2. Using only images without artifacts and hair, being easy or medium to classify;
3. Using only images without artifacts but with hair, being easy or medium to classify.

Experimental design: the Atlas dataset was randomly divided in 10 folds balanced in terms of melanoma/non-melanoma images, types of skin lesion and difficulty. Several images had a black frame that were previously removed. The images were all scaled to 100,000 pixels each one, since they had different sizes. The dataset contains clinical and dermoscopic images of each lesion: they were both used in the experiments but the fold division was done in terms of *cases*, not images, to eliminate risk of contamination. The BossaNova parametrization was the one that led to better results for Spatial Pyramids Match (SPM), since it was the spatial pooling strategy employed in this experiment (scales: 1×1 and 2×2). The parameter values are: codebook size ($M = 2048$), number of bins ($B = 4$), and alphas values limited between $\alpha_{MIN} = 0.6$ and $\alpha_{MAX} = 1.6$.

Results: the average AUC for each sub-experiment was:

1. Complete dataset removing difficult lesions: 80.0%
2. Without artifacts and hair (easy/medium lesions): 85.0%
3. Without artifacts but *with* hair (easy/medium lesions): 85.0%

We also performed an extra sub-experiment to identify if the clinical images were impacting the results. We used the same folds of experiment (1), just removing the clinical images. The AUC of this setup achieved 88.0%.

Analysis: the results achieved indicate that the Atlas is a very challenging dataset, because our previous result of 93.7% decreased to 80.0%. It can be explained by the fact that here we are using both clinical and dermoscopic images in the experiments. This hypothesis is proven by the extra validation using the same folds but eliminating the clinical images, which leads to an AUC of 88.0%. This observation may suggest that even advanced approaches for automated melanoma screening are not prepared to deal with clinical images captured by common cameras, since clinical and dermoscopic images deal with different aspects of the problem and may not be mixed in the experiments. Another factor that could have mislead the classifier is the presence of artifacts in the images (like rules, dots, hair and arrows). This is confirmed by the second experiment in which

we didn't use images with artifacts, improving the AUC to 85.0% even with clinical images in the folds. Although we were expecting a higher improvement, we concluded that filtering the datasets is essential for the classification methods and our approach is not fully robust to noise in images. However, the third experiment (images without artifacts but with hair) also achieved an AUC of 85.0%, suggesting that our method is, indeed, robust to hair. Since in a medical examination the lesion images can be easily obtained without artifacts but the hair removal is not straightforward, we conclude that our approach can be a powerful tool for automated melanoma screening.

These experiments show that our approach is robust to hair in skin lesion images. It was shown that clinical images are, for the moment, very challenging for automated melanoma screening since they introduce some difficulties, like brightness and lack of details, that can mislead the classifier. Despite difficulties, this kind of image must be used for screening purposes since it is easier and cheaper to obtain. When the experiment is done without clinical images, even keeping the artifacts (e.g. hair), the AUC is almost the same than the previous ones. But, surprising, the training set size of this experiment is bigger than the ones done with IRMA dataset. This indicates that in the Atlas dataset our solution didn't perform as well as before, reinforcing two aspects: (1) this dataset is really challenging and, (2) there is space for more improvements in our method. This will be discussed next.

4.3.6 A Critical Review of Our Benchmark

A comparison with other methods which use the same dataset was required to investigate whether the results not so satisfactory of the previous experiment were caused by problems in our method or intrinsic difficulties of the images. We choose the work of Wadhawan et al. [2011a] as benchmark since it is one of the most detailed methods in literature that employs a BoVW model for melanoma classification. The choice was particularly interesting because while our result was 85.0%, the authors reported an AUC of 91.1% even using poor image feature descriptors and a simple BoVW pipeline. This section describes how [Wadhawan et al., 2011a] was reproduced and how we compare to it. For a positioning of the selected job front of the literature, see Chapter 2.

Goal: to make a direct comparison between our approach with other method of the literature using the same dataset in order to have a better idea of the classification power of our solution to automated melanoma screening.

Experimental design: the benchmark was implemented accordingly to the directions presented in the original paper. Some parameters were not fully described, so we have

contacted the authors: we had no answer. When we contacted the authors again, they refused to share detailed information claiming intellectual property restrictions since the method was in a patenting process. So, for these cases we applied typical values found in literature. We first used the images without artifact and hair from Atlas dataset, but we included the images regardless their difficulty (using the easy, medium and difficult ones). The selected images were divided in a 10-fold cross-validation schema.

Results: the result registered an AUC of just 75.9%, very far from the 91.1% reported in the original paper.

Analysis: despite of the best effort to reproduce the original results of [Wadhawan et al., 2011a], our attempts didn't achieve what was published. Remembering that this is one of the most well detailed papers about automated melanoma screening, this indicates a serious problem of the literature: the results are **not** reproducible. After removing the difficult images of the original selection and running the experiment again, the AUC achieved the same value as we did previously (85.0%). But, it is very important to highlight that the parametrization for this experiment was refined. So we believe that we achieved the best result that this method could generate. However, our approach was validated in the Atlas dataset using the parametrization especially designed for the IRMA dataset. It is also important to note that in this experiment the images were used in their original size, but we have re-sized them to validate our approach. This could has limited our feature extraction step. All of these observations are relevant to argue that our approach results for the Atlas dataset can still be improved.

The strongest conclusion of this experiment is that the literature of automated melanoma screening has critical problems that must be resolved. The lack of details in the papers is the biggest obstacle for reproducibility. Since there is no public dataset of skin lesion, reproducibility is essential for fair comparison between new approaches and existing ones. This experiment reveals, yet, that our results are very promising and there is space for further improvements in our approach. The enhanced mid-level representations, as expected, perform better results than traditional BoVW implementations.

4.4 Conclusion

This chapter described our experiments. First and foremost, we proved that BossaNova, an enhanced mid-level representation, performs better than the traditional BoVW implementation over skin lesion classification. Then, we analyzed which step of the classification

pipeline is more relevant for the improvement of the classifier, and also explored a BossaNova parametrization for the IRMA dataset.

Other important experiment was the validation of the proposed SCP, a novel spatial pooling strategy especially designed for automated melanoma screening. Despite of the expectations and medical support for the model, the approach was not so effective for the problem but gave us some insights for further inspections.

The third great contribution of this work was an analysis of the literature reproducibility. It was shown that it is not possible to reproduce the current state-of-the-art due to the lack of information in the published papers. It is also important to note that none of the state-of-the-art works reported standard deviation of their results, preventing a deeper analysis of the actual behavior of the literature. We, however, presented our deviations by showing the residuals of the ANOVA analysis. We hope that this work can motivate other researchers to make their methods easy to be compared and reproduced.

5 Conclusions

This chapter summarizes the main findings and points out future guidelines of this work. Although the main objectives were satisfied, we left some areas of interest to be explored in the future due to time and scope limitations.

5.1 Main Findings

Our experiments were very important to elucidate open questions of automated melanoma screening. The experiments were detailed in Chapter 4, in which we also provided an analysis of the results. Here we recapitulate the main findings and contributions organized on four aspects:

- **Problem comprehension**: this work provided an analytical and critical revision of the automated melanoma classification literature. We focused in solutions based in the BoVW model since it is the image classification approach more indicated for this kind of problem, that is, classification of specific small datasets. We discovered that the low-level feature extraction is the most important step for the classification, so authors should pay attention to the feature descriptors employed in their systems. We also proved that the classifiers tend to improve when the number of samples in the training phase is bigger, indicating that melanoma datasets should be made available in order to enable deeper investigations and methods enhancement. Finally, we also showed that clinical images are very difficult to be applied for automated melanoma screening, since they do not present skin lesion details and may mislead the classifier. Nevertheless this kind of data is very important to enable melanoma screening with mobile devices when the image acquisition may not be perfectly controlled.

- **Introduction of new techniques**: this work is particularly important for the melanoma screening community that uses BoVW-based approaches in their solutions. We claim this because we proved that modern extensions of the BoVW model, that introduce enhanced mid-level representations, are more adequate for image classification, leading to better results. Our experiments were based using the BossaNova mid-level descriptors that overcome the classification rates of the traditional BoVW implementation in several scenarios, including the melanoma one. Our approach was able to generate a melanoma classifier with AUC up to 93.7% in a controlled dataset (IRMA

dataset) and up to 88.0% in a more challenging one using only dermoscopic images (Atlas dataset). Other experiments demonstrated that our solution is partially robust to noise in the images (like dots, rulers and pointers) but fully robust to the presence of hair, which is desired by physicians. This work also introduced a spatial pooling strategy (SCP) especially designed for melanoma image classification. Despite of its medical theoretical basis, the experiments showed that the proposed pooling is not so effective to the problem, but indicates that asymmetry is one of the major criteria to be investigated in the lesions to identify melanomas although can not be interpreted as a ground truth since some melanomas are symmetric. Nevertheless, we hope that future improvements in the SCP can lead to better results.

- **Reproducibility**: other relevant aspect of this research is that we demonstrate that the automated melanoma screening literature faces critical problems of reproducibility. When we tried to reproduce one of the most detailed works, we were not able to achieve the reported results. In this sense, our contributions are to draw attention to this fact, to employ rigorous evaluation protocols to the methods proposed and to make our code freely available in order to stimulate other authors to do the same.

- **Portability to mobile environment**: finally, although melanoma screening using mobile devices is not our main goal, this work opens the opportunity for future investigations in this area. Since mobile devices will generate clinical images of the skin lesions, the previous findings of this work give an intuition for the difficulty of the problem. Also, compact feature descriptors must be explored to enable data processing in limited resources environments.

We also would like to highlight that the results achieved in this work are very promising and suggest an advance in the state-of-the-art of the automated melanoma screening problem. Our efforts so far have culminated in the publication of the conference paper[1]:

***Fornaciali, M.**, Avila, S., Carvalho, M., & Valle, E. (2014). Statistical Learning Approach for Robust Melanoma Screening. In Proceedings of the 2014 27th SIBGRAPI Conference on Graphics, Patterns and Images (pp. 319-326). IEEE Computer Society.*

5.2 Future Work

The experiments done so far gave us some insights to go further into the research. We found that other challenges on melanoma screening problem, such as the incorporation of

[1] The code and result details of this paper are publicly available - https://sites.google.com/site/robustmelanomascreening/

histopathologic data and cross-dataset experiments are unpublished in literature.

We are also interested in investigating the human ability to classify a skin lesion, particularly, the percentage of agreement among physicians to annotate an lesion as malignant or not. This kind of result is important because if it is lower than our results, this would prove that our melanoma classifier could support a physician in classifying a skin lesion but, of course, not being able to replace him/her.

We can improve the framework to increase the melanoma screening rate, since we have not exhausted the experiments with the proposed SCP. Improvements in SCP include, but are not limited to, new image divisions in order to better and easily identify asymmetry of the lesion without the need of segmenting it.

Extensions of this research focuses on improving the overall framework. This can be done working on improvements in both basic steps of the BoVW model: the low-level and the mid-level feature extraction.

Regarding the low-level feature extraction, we can enrich it by incorporating learning steps based on Deep Learning Architectures. This can improve the informative power of the data that feed the mid-level. Compact descriptors that keep valuable information would be also appreciated to accelerate the classification without loss of accuracy. We can also explore normalization methods that transform low-level features (like SIFT) into powerful informative data, for example, employing the method proposed by Kobayashi [2013], a new alternative for PCA.

For the mid-level, we aim to experiment new pooling schemes that add more information to the classifier. Alternatives include, but are not limited to, the "generalized max pooling" proposed by Murray and Perronnin [2014], that improves the pooling schema specially for Fisher Vector [Perronnin et al., 2010]. Our interest in Fisher Vector relies on its good complementarity with BossaNova, the base of our framework. Other new pooling approach to be investigated is one proposed by Fanello et al. [2013]. They provided an extension of the standard SPM representation that can also favor our SCP strategy.

On the other hand, CVPR[2] 2015 shows that advances in DLA have demonstrated competitive results with the traditional BoVW extensions. Researches in both fields suggest that the techniques could be combined since they present complementary advantages, as pointed by Perronnin and Larlus [2015], that combined Fisher Vector with Deep Learning. Also related with DLA, other approach to be considered is the transfer learning methods, which surprisingly uses information trained in a dataset to classify images of other scope,

[2] CVPR: Conference on Computer Vision and Pattern Recognition. This is the most important computer vision conference.

showing interesting results.

As pointed in Chapter 2, melanoma screening can also be done in mobile environment through handheld devices. Since we achieved state-of-the-art results, it would be an important contribution to carry our framework to mobile devices, paying attention to the fact that our solution should be redesigned to better deal with clinical images. This leads to a number of space and memory limitations that can negatively affect the classification. In order to overcome these issues, the extension of this work can investigate compact representations for image classification, already proposed by Zhang et al. [2014]. We aim to analyze if these representation for general-purpose image classification is also suitable for special-purpose image classification.

5.3 Final Remarks

Concluding this work, we would like to reinforce that, despite of the interests and importance of automated screening, no system, with the current techniques and knowledge, can replace the opinion of a physician. Nevertheless, automated screening is a powerful tool to optimize time and cost.

Due to the promising results of this research, it can be extended in order to provide screening methods for other diseases, besides of melanoma. Other fields of Medicine that can benefit from this work, just to cite a few examples, are cardiology (studying echocardiography images in order to identify heart diseases), neurology (classification of regions of interest and types of brain lesions), ophthalmology (diabetic retinopathy detection), oncology (processing mammography images in order to classify breast cancer) and others.

Bibliography

American Cancer Society. Cancer Facts & Figures 2013, 2013. Last accessed: July 31, 2015. Cited in page 1.

International Skin Imaging Collaboration, 2015. http://www.isdis.net/. Last accessed: June 06, 2015. Cited in page 21.

Q. Abbas, M. Celebi, I. Garcia, and W. Ahmad. Melanoma recognition framework based on expert definition of ABCD for dermoscopic images. *Skin Research and Technology*, 19(1): 93–102, 2012. Cited 5 times in pages 15, 17, 19, 20, and 42.

M. Abedini, N. C. F. Codella, J. H. Connell, R. Garnavi, M. Merler, S. Pankanti, J. R. Smith, and T. Syeda-Mahmood. A generalized framework for medical image classification and recognition. *IBM Journal of Research and Development*, 59(2/3):1–1, 2015. Cited 3 times in pages 1, 19, and 20.

O. Abuzaghleh, B. D. Barkana, and M. Faezipour. Automated skin lesion analysis based on color and shape geometry feature set for melanoma early detection and prevention. In *Systems, Applications and Technology Conference (LISAT), 2014 IEEE Long Island*, pages 1–6. IEEE, 2014. Cited 2 times in pages 19 and 20.

Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid. Good practice in large-scale learning for image classification. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(3):507–520, 2014. Cited in page 12.

S. Allen. Melanoma Screening Saves Lives, 2015. http://www.skincancer.org/skin-cancer-information/melanoma/melanoma-prevention-guidelines/melanoma-screening-saves-lives. Last accessed: June 06, 2015. Cited in page 21.

R. Arandjelovic and A. Zisserman. Three things everyone should know to improve object retrieval. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 2911–2918, 2012. Cited 3 times in pages 27, 28, and 38.

R. Arandjelovic and A. Zisserman. All about vlad. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 1578–1585, 2013. Cited in page 13.

G. Argenziano, G. Fabbrocini, P. Carli, V. De Giorgi, E. Sammarco, and M. 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. *Arch Dermatology*, 134(12):1563–1570, 1998. Cited 4 times in pages 15, 16, 18, and 31.

G. Argenziano, H. P. Soyer, V. De Giorgi, D. Piccolo, P. Carli, M. Delfino, et al. Dermoscopy: a tutorial. *EDRA, Medical Publishing & New Media*, 2002. Cited in page 16.

G. Argenziano, C. Catricalà, M. Ardigo, P. Buccini, P. De Simone, L. Eibenschutz, A. Ferrari, G. Mariani, V. Silipo, I. Sperduti, et al. Seven-point checklist of dermoscopy revisited. *British Journal of Dermatology*, 164(4):785–790, 2011. Cited 2 times in pages 17 and 31.

S. Avila. *Extended bag-of-words formalism for image classification*. PhD thesis, Federal University of Minas Gerais & Pierre and Marie Curie University, 2013. Cited in page 11.

S. Avila, N. Thome, M. Cord, E. Valle, and A. de A. Araújo. BossaNova at ImageCLEF 2012 Flickr Photo Annotation Task. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF)*, 2012. Cited in page 37.

S. Avila, N. Thome, M. Cord, E. Valle, and A. de A. Araújo. Pooling in image representation: the visual codeword point of view. *Computer Vision and Image Understanding (CVIU)*, 117(5):453–465, 2013. Cited 9 times in pages 6, 9, 12, 14, 23, 26, 29, 30, and 37.

R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley, 1st edition, 1999. Cited in page 6.

C. Barata, J. Marques, and T. Mendonça. Bag-of-features classification model for the diagnose of melanoma in dermoscopy images using color and texture descriptors. In *International Conference on Image Analysis and Recognition (ICIAR)*, pages 547–555, 2013. Cited 2 times in pages 18 and 20.

C. Barata, M. Ruela, M. Francisco, T. Mendonça, and J. S. Marques. Two systems for the detection of melanomas in dermoscopy images using texture and color features. *Systems Journal, IEEE*, 8(3):965–979, 2014. Cited 2 times in pages 19 and 20.

A. Bastawrous. Portable Eye Examination Kit (Peek), 2012. http://www.peekvision.org/. Last accessed: July 31, 2015. Cited in page 2.

H. Bay, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. In *Computer Vision–ECCV 2006*, pages 404–417. Springer, 2006. Cited in page 8.

J. Benois-Pineau, F. Precioso, and M. Cord. *Visual indexing and retrieval*. Springer Verlag, 2012. Cited 2 times in pages 9 and 11.

M. Bento, L. Rittner, S. Appenzeller, A. Lapa, and R. Lotufo. Analysis of brain white matter hyperintensities using pattern recognition techniques. In *SPIE Medical Imaging*. International Society for Optics and Photonics, 2013. Cited in page 5.

Y. Boureau, F. Bach, Y. LeCun, and J. Ponce. Learning mid-level features for recognition. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 2559–2566, 2010. Cited 3 times in pages 9, 18, and 29.

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*, 32(16):2187–2196, 2011. Cited 3 times in pages 15, 19, and 31.

G. Castellano, R. Lotufo, L. Bonilla, L. M. Li, and F. Cendes. Processamento de imagens de ressonância magnética tridimensional em neurologia: vantagens e dificuldades. *Rev. Bras. Neurol*, 39(3):5–14, 2003. Cited in page 5.

M. E. Celebi, G. Schaefer, H. Iyatomi, W. V. Stoecker, J. M. Malters, and J. M. Grichnik. An improved objective evaluation measure for border detection in dermoscopy images. *Skin Research and Technology*, 15(4):444–450, 2009. Cited in page 31.

C.-C. Chang and C.-J. Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011. Cited in page 36.

K. Chatfield, V. S. Lempitsky, A. Vedaldi, and A. Zisserman. The devil is in the details: an evaluation of recent feature encoding methods. In *BMVC*, volume 2, page 8, 2011. Cited in page 7.

G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision, ECCV*, volume 1, pages 1–2, 2004. Cited in page 6.

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 *43rd Hawaii International Conference on System Sciences (HICSS)*, pages 1–10, 2010. Cited 2 times in pages 19 and 31.

C. Doukas, P. Stagkopoulos, C. Kiranoudis, and I. Maglogiannis. Automated skin lesion assessment using mobile technologies and cloud platforms. In *International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 2444–2447, 2012. Cited 4 times in pages 18, 19, 20, and 21.

S. Fanello, N. Noceti, C. Ciliberto, G. Metta, and F. Odone. Ask the image: Supervised pooling to preserve feature locality. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 851–858, 2013. Cited in page 51.

O. Faust, R. Acharya, E. Y.-K. Ng, K.-H. Ng, and J. S. Suri. Algorithms for the automated detection of diabetic retinopathy using digital fundus images: a review. *Journal of Medical Systems*, 36(1):145–157, 2012. Cited in page 1.

A. Fidalgo Barata, E. Celebi, and J. Marques. Improving dermoscopy image classification using color constancy. 2014. Cited 2 times in pages 19 and 20.

M. Fornaciali, S. Avila, M. Carvalho, and E. Valle. Statistical learning approach for robust melanoma screening. In *27th SIBGRAPI Conference on Graphics, Patterns and Images*, pages 319–326. IEEE Computer Society, 2014. Cited 2 times in pages 20 and 40.

K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *Computer Vision–ECCV 2014*, pages 346–361. Springer, 2014. Cited in page 12.

G. Hinton. Deep belief networks. In *Scholarpedia 4 (5): 5947. doi:10.4249/scholarpedia.5947*, 2009. Cited in page 6.

Y. Huang, Z. Wu, L. Wang, and T. Tan. Feature coding in image classification: A comprehensive study. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(3): 493–506, 2014. Cited in page 14.

H. Iyatomi, H. Oka, M. 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*, 32(7):566–579, 2008. Cited 4 times in pages 15, 19, 20, and 42.

R. Jain. *The Art of Computer Systems Performance Analysis: techniques for experimental design, measurement, simulation, and modeling*. John Wiley & Sons, Inc., 1991. Cited 2 times in pages 37 and 38.

H. Jégou, M. Douze, C. Schmid, and P. Pérez. Aggregating local descriptors into a compact image representation. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 3304–3311, 2010. Cited 2 times in pages 12 and 13.

A. P. Kassianos, J. D. Emery, P. Murchie, and F. M. Walter. Smartphone applications for melanoma detection by community, patient and generalist clinician users: a review. *British Journal of Dermatology*, 2015. Cited in page 21.

B. Klein, G. Lev, G. Sadeh, and L. Wolf. Associating neural word embeddings with deep image representations using fisher vectors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4437–4446, 2015. Cited in page 14.

T. Kobayashi. Dirichlet-based histogram feature transform for image classification. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 3278–3285, 2013. Cited in page 51.

P. Koniusz, F. Yan, and K. Mikolajczyk. Comparison of mid-level feature coding approaches and pooling strategies in visual concept detection. *Computer Vision and Image Understanding*, 117(5):479–492, 2013. Cited in page 9.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. Cited in page 6.

S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 2169–2178, 2006. Cited 8 times in pages 12, 13, 14, 30, 33, 39, 40, and 41.

J. Li and N. M. Allinson. A comprehensive review of current local features for computer vision. *Neurocomputing*, 71(10):1771–1787, 2008. Cited in page 8.

P. Li, X. Lu, and Q. Wang. From dictionary of visual words to subspaces: Locality-constrained affine subspace coding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2348–2357, 2015. Cited 2 times in pages 12 and 14.

Y. Li, L. Liu, C. Shen, and A. van den Hengel. Mid-level deep pattern mining. *arXiv preprint arXiv:1411.6382*, 2014. Cited in page 14.

D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision (IJCV)*, 60:91–110, 2004. Cited 2 times in pages 6 and 8.

J. March, M. Hand, and D. Grossman. Practical application of new technologies for melanoma diagnosis: Part i. noninvasive approaches. *Journal of the American Academy of Dermatology*, 72(6):929–941, 2015. Cited in page 21.

J. Marques, C. Barata, and T. Mendonca. On the role of texture and color in the classification of dermoscopy images. In *International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 4402–4405, 2012. Cited 3 times in pages 18, 19, and 20.

W. S. McCulloch and W. Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of Mathematical Biophysics*, 5(4):115–133, 1943. Cited in page 6.

T. Mendonca, P. M. Ferreira, J. S. Marques, AR. S. Marcal, and J. Rozeira. Ph2 - a dermoscopic image database for research and benchmarking. In *Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE*, pages 5437–5440, July 2013. doi: 10.1109/EMBC.2013.6610779. Cited in page 31.

M. Mete and N. Sirakov. Dermoscopic diagnosis of melanoma in a 4d space constructed by active contour extracted features. *Computerized Medical Imaging and Graphics*, 36(7): 572–579, 2012. Cited 2 times in pages 15 and 19.

K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 27(10):1615–1630, 2005. Cited in page 8.

E. Mikos, I. Sioulas, K. Sidiropoulos, and D. Cavouras. An android-based pattern recognition application for the characterization of epidermal melanoma. In *Workshop on Biomedical Instrumentation and Related Engineering and Physical Sciences (BIOMEP)*, 2012. Cited 3 times in pages 19, 20, and 21.

N. Murray and F. Perronnin. Generalized max pooling. *IEEE Computer Vision and Pattern Recognition (CVPR)*, 2014. Cited in page 51.

F. Nachbar, W. Stolz, Tanja. Merkle, A. Cognetta, T. Vogt, M. Landthaler, P. Bilek, O. Braun-Falco, and G. Plewig. The ABCD rule of dermatoscopy: High prospective value in the diagnosis of doubtful melanocytic skin lesions. *Journal of the American Academy of Dermatology*, 30(4):551–559, 1994. ISSN 0190-9622. Cited 3 times in pages 15, 31, and 32.

J. H. Neltner, E. L. Abner, F. A. Schmitt, S. K. Denison, S. Anderson, E. Patel, and P. T. Nelson. Digital pathology and image analysis for robust high-throughput quantitative assessment of alzheimer disease neuropathologic changes. *Journal of Neuropathology and Experimental neurology*, 71(12):1075, 2012. Cited in page 1.

F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorization. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2007. Cited 2 times in pages 12 and 13.

F. Perronnin and D. Larlus. Fisher vectors meet neural networks: A hybrid classification architecture. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 3743–3752, 2015. Cited in page 51.

F. Perronnin, J. Sánchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. In *Computer Vision–ECCV 2010*, pages 143–156. Springer, 2010. Cited 5 times in pages 6, 12, 13, 30, and 51.

D. Picard and P-H. Gosselin. Improving image similarity with vectors of locally aggregated tensors. In *IEEE International Conference on Image Processing (ICIP)*, pages 669–672, 2011. Cited 2 times in pages 12 and 13.

R. Pires, S. Avila, H. F. Jelinek, J. Wainer, E. Valle, and A. Rocha. Automatic diabetic retinopathy detection using bossanova representation. In *International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2014a. Cited in page 5.

R. Pires, H. F. Jelinek, J. Wainer, E. Valle, and A. Rocha. Advancing bag-of-visual-words representations for lesion classification in retinal images. *PloS one*, 9(6):e96814, 2014b. Cited in page 5.

D. S. Rigel. Epidemiology of melanoma. In *Seminars in cutaneous medicine and surgery*, volume 29, pages 204–209. WB Saunders, 2010. Cited 2 times in pages 1 and 15.

L. Rittner, S. Appenzeller, and R. Lotufo. Segmentation of brain structures by watershed transform on tensorial morphological gradient of diffusion tensor imaging. In *Computer Graphics and Image Processing (SIBGRAPI), 2009 XXII Brazilian Symposium on*, pages 126–132. IEEE, 2009. Cited in page 5.

J. M. Rondina, R. de A. Lotufo, and M. A. Gutierrez. Um sistema de segmentação interativa do ventrículo esquerdo em seqüências de imagens de ressonância magnética (cine mr). *Rev. Bras. Eng. Biomed*, 18(3):117–131, 2002. Cited in page 5.

SCF. Skin Cancer Foundation, 2013. http://www.skincancer.org/. Last accessed: July 31, 2015. Cited 2 times in pages 1 and 17.

J. Scharcanski and M. E. Celebi. *Computer vision techniques for the diagnosis of skin cancer*. Springer, 2014. Cited in page 19.

N. Situ, X. Yuan, J. Chen, and G. Zouridakis. Malignant melanoma detection by bag-of-features classification. In *International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 3110–3113, 2008. Cited 4 times in pages 18, 19, 20, and 42.

J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *International Conference on Computer Vision (ICCV)*, volume 2, 2003. Cited in page 6.

P. Skaane, A. I. Bandos, R. Gullien, E. B. Eben, U. Ekseth, U. Haakenaasen, M. Izadi, I. N. Jebsen, G. Jahr, M. Krager, et al. Comparison of digital mammography alone and digital mammography plus tomosynthesis in a population-based screening program. *Radiology*, 267(1):47–56, 2013. Cited in page 1.

A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 22(12):1349–1380, 2000. Cited in page 6.

H. P. Soyer, G. Argenziano, I. Zalaudek, R. Corona, F. Sera, R. Talamini, F. Barbato, A. Baroni, L. Cicale, A. Di Stefani, et al. Three-point checklist of dermoscopy. *Dermatology*, 208(1):27–31, 2004. Cited 2 times in pages 17 and 31.

T. Tuytelaars. Dense interest points. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 2281–2288, 2010. Cited in page 28.

T. Tuytelaars and K. Mikolajczyk. Local invariant feature detectors: a survey. *Foundations and Trends® in Computer Graphics and Vision*, 3(3):177–280, 2008. Cited in page 8.

A. Tyagi, K. Miller, and M. Cockburn. e-health tools for targeting and improving melanoma screening: a review. *Journal of Skin Cancer*, 2012, 2012. Cited in page 22.

J. van Gemert, C. Veenman, A. Smeulders, and J-M. Geusebroek. Visual word ambiguity. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 32:1271–1283, 2010. Cited in page 9.

V. N. Vapnik. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., 1995. Cited 2 times in pages 8 and 19.

T. Wadhawan, N. Situ, K. Lancaster, X. Yuan, and G. Zouridakis. Skinscan©: A portable library for melanoma detection on handheld devices. In *IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 133–136, 2011a. Cited 7 times in pages 18, 19, 20, 21, 42, 46, and 47.

T. Wadhawan, N. Situ, H. Rui, K. Lancaster, X. Yuan, and G. Zouridakis. Implementation of the 7-point checklist for melanoma detection on smart handheld devices. In *International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 3180–3183, 2011b. Cited 5 times in pages 15, 18, 19, 20, and 21.

J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3360–3367. IEEE, 2010. Cited in page 14.

J. A. Wolf, J. F. Moreau, O. Akilov, T. Patton, J. C. English, J. Ho, and L. K. Ferris. Diagnostic inaccuracy of smartphone applications for melanoma detection. *JAMA Dermatology*, 149(4):422–426, 2013. Cited 3 times in pages 3, 21, and 22.

J. Yang, K. Yu, Y. Gong, and T. Huang. Linear spatial pyramid matching using sparse coding for image classification. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 1794–1801, 2009. Cited in page 29.

Y. Zhang, J. Wu, and J. Cai. Compact representation for image classification: To choose or to compress? In *IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 907–914, 2014. Cited in page 52.

X. Zhou, K. Yu, T. Zhang, and T. S. Huang. Image classification using super-vector coding of local image descriptors. In *Computer Vision–ECCV 2010*, pages 141–154. Springer, 2010. Cited 2 times in pages 12 and 13.

G. Zouridakis, T. Wadhawan, N. Situ, R. Hu, X. Yuan, K. Lancaster, and C. M. Queen. Melanoma and other skin lesion detection using smart handheld devices. *Mobile Health Technologies: Methods and Protocols*, pages 459–496, 2015. Cited in page 21.