High-Level Intuitive Features (HLIFs) for Melanoma Detection

by

Robert Amelard

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Applied Science
in
Systems Design Engineering

Waterloo, Ontario, Canada, 2013

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Feature extraction of segmented skin lesions is a pivotal step for implementing accurate decision support systems. Existing feature sets combine many ad-hoc calculations and are unable to easily provide intuitive diagnostic reasoning. This thesis presents the design and evaluation of a set of features for objectively detecting melanoma in an intuitive and accurate manner. We call these “high-level intuitive features” (HLIFs).

The current clinical standard for detecting melanoma, the deadliest form of skin cancer, is visual inspection of the skin’s surface. A widely adopted rule for detecting melanoma is the “ABCD” rule, whereby the doctor identifies the presence of **A**symmetry, **B**order irregularity, **C**olour patterns, and **D**iameter. The adoption of specialized medical devices for this cause is extremely slow due to the added temporal and financial burden. Therefore, recent research efforts have focused on detection support systems that analyse images acquired with standard consumer-grade camera images of skin lesions. The central benefit of these systems is the provision of technology with low barriers to adoption. Recently proposed skin lesion feature sets have been large sets of low-level features attempting to model the widely adopted ABCD criteria of melanoma. These result in high-dimensional feature spaces, which are computationally expensive and sparse due to the lack of available clinical data. It is difficult to convey diagnostic rationale using these feature sets due to their inherent ad-hoc mathematical nature.

This thesis presents and applies a generic framework for designing HLIFs for decision support systems relying on intuitive observations. By definition, a HLIF is designed explicitly to model a human-observable characteristic such that the feature score can be intuited by the user. Thus, along with the classification label, visual rationale can be provided to further support the prediction. This thesis applies the HLIF framework to design 10 HLIFs for skin cancer detection, following the ABCD rule. That is, HLIFs modeling asymmetry, border irregularity, and colour patterns are presented.

This thesis evaluates the effectiveness of HLIFs in a standard classification setting. Using publicly-available images obtained in unconstrained environments, the set of HLIFs is compared with and against a recently published low-level feature set. Since the focus is on evaluating the features, illumination correction and manually-defined segmentations are used, along with a linear classification scheme. The promising results indicate that HLIFs capture more relevant information than low-level features, and that concatenating the HLIFs to the low-level feature set results in improved accuracy metrics. Visual intuitive information is provided to indicate the ability of providing intuitive diagnostic reasoning to the user.

Acknowledgements

"Acknowledgement" is not expressive enough to describe my gratitude to my supervisors, Alex Wong and David Clausi. Your guidance and rock-solid support has helped me discover latent passions and has opened my eyes to new ways of thinking, both professionally and personally. I am very fortunate to have had you both as supervisors, both in the roles of inspirational individuals and as a supportive team. Looking back at the past two years, I feel like I am giant leap ahead from when I started.

A big "thank you!" to my mentor and advocate, George Labahn. Thank you for introducing me to the wild world of research! It is blindingly clear to me how my terms under your supervision have platformed me for success in the research world.

Thank you to the VIP lab members, who are always willing to brainstorm and critically analyse my work and thoughts. The VIP lab has an uncanny recipe to foster impactful work through support and enthusiasm.

Thanks as well to my readers, Andrea Scott and Paul Fieguth. I greatly appreciate the time and effort you have put in to help me produce a sound thesis. It is in good hands.

Finally, I would like to acknowledge National Natural Sciences and Engineering Research Council of Canada (NSERC), Agfa Healthcare, and Ontario Centres of Excellence (OCE) for funding support.

Dedication

To Kaylen – for helping me see the forest, challenging me to progress beyond plateaus, and sharing my excitement in matters both big and small.

To Richard, Mom, Dad – my unyielding supporters from childhood endeavours through academic pursuits.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

This thesis presents a set of intuitive features for describing the malignant nature of segmented skin lesion images obtained using consumer-grade cameras. A general framework for designing high-level intuitive features (HLIFs) from which the diagnostic rationale can be relayed to the doctor in an intuitive manner is also presented. The proposed set of skin lesion features follows this framework.

1.1 What Is Melanoma?

Melanoma is the deadliest form of skin cancer [1]. In a report published in 2000, the World Health Organization (WHO) estimated that approximately 65,000 global deaths related to melanoma occurred that year [2]. If caught early, a simple extraction of the cancerous tissue can completely cure the patient of melanoma. However, if identified late, the cancer can spread and the prognosis is bleak.

In North America, melanoma cases are typically identified by a dermatologist or by a pathologist (given a biopsy by the dermatologist). Referral to a dermatologist is usually issued by a general practitioner from a visual identification of an abnormal lesion. Furthermore, dermatologists usually employ naked-eye examination for determining malignant lesions, sometimes with the aid of a dermatoscope [3]. A dermatoscope is an optical device used by some dermatologists to magnify and enhance skin structure while mitigating skin surface reflection. In either case, melanoma detection is performed according to a visual examination.

Unfortunately, several factors make it difficult to visually identify melanoma. During its early to mid stages, melanoma can be very similar in appearance to benign dysplastic nevi (i.e., “moles”) to the naked eye. Melanoma can also be observed in many shapes and forms, making it even more difficult to correctly identify a new lesion as a melanoma. As we will see in Chapter 2, identifying melanoma as early as possible is crucial to patient prognosis. Furthermore, although it causes 75% of all skin-cancer related deaths [2], only 5% of skin cancer cases are reported as melanoma [4]. Paired with its varying appearances, many of which look similar to a benign dysplastic nevus (i.e., “mole”), this low incidence rate makes it challenging to precisely identify all melanoma cases while maintaining a realistic true negative (i.e., benign) diagnostic rate.

1.2 What Is Being Done?

There has been a number of decision support systems proposed in the literature. The typical “black box” model of these systems is to receive an image of a skin lesion and output whether the lesion is suspected to be malignant or benign. An important part of this analysis (and consequently the main topic of this thesis) is *feature extraction*, whereby predetermined calculations are performed using the image to represent the visual information as a set of real numbers. This set is often called the *feature vector*. The performance of the classification is highly dependent on the success of feature extraction.

1.2.1 Problem 1: Dermoscopic Images

Most of the research literature proposes methods of identifying melanoma given a dermoscopic image that has been obtained using a digital dermatoscope. These devices produce very “clean” images that are clear of skin surface reflection and elucidate skin surface structure. However, there are two inherent issues with these systems. First, a recent survey has found that only 48% of respondents from fellows of the American Academy of Dermatology reported that they use dermatoscopes to assess skin lesions [3]. It was not specified how many were digital dermatoscope users (i.e., use a dermatoscope that can capture an image), however it can be postulated that fewer still use such devices. These findings limit the scope of methods that use such images. Second, methods using dermoscopy images are restricted to professional dermatology settings, as patients and general practitioners are unlikely to possess such a specialized device.

1.2.2 Problem 2: Low-Level Features

The majority of existing work has focused on large sets of low-level features. These are calculations that have not been designed to model a particular observable characteristic, but rather are used en masse in an attempt to capture some high-level characteristic. This trend may be due to the focus on preprocessing and segmentation techniques rather than feature extraction techniques. These highly-dimensional feature spaces become especially problematic when working with the small data sets that are available to the scientific community. Overfitting and the lack of statistical validity are major concerns in sparse feature spaces.

1.2.3 Problem 3: Gaining the Doctor's Trust

To the best of the author's knowledge, the output of existing skin cancer decision support systems is either a binary or probabilistic classification of malignancy. However, a single label may not be enough to gain the trust of the doctor using the system, especially given the critical nature of the problem. No matter how sophisticated or accurate a system may be, it may be rejected in the absence of user trust [5]. System credibility has received significant attention in human-computer interaction research [6]; however these ideals have not been explicitly introduced to melanoma decision support system research.

1.3 Solution – Thesis Contributions

This thesis presents a set of high-level intuitive features (HLIFs) which can capture specific high-level information and provide intuitive diagnostic rationale to the user with the goal of increasing system credibility and strengthening the user-system trust. The design rationale of each feature is given so that the feature can be easily interpreted and relayed to the doctor in a dermatology setting. Ideally, a decision support system incorporating these HLIFs can be issued to patients, general practitioners, and dermatologists who do not use a digital dermatoscope in their clinical setup. The barrier to adoption is minimal, since standard camera images are analysed. For the system to be valid under the unconstrained image acquisition environments, the features have been designed to be resistant to changes in resolution, camera quality, and lighting conditions.

1.4 Thesis Outline

The remainder of this thesis is organized as follows. Chapter 2 provides relevant background information needed to understand the relevancy of this thesis' contribution. Chapter 3 presents the framework for designing HLIFs, and Chapter 4 presents the design and calculation of 10 HLIFs for describing skin lesion images using this framework. Chapter 5 provides experimental results and discussions by combining and comparing the proposed features with a state-of-the-art low-level feature set using standard consumer-grade camera images. Conclusions are drawn and recommended future work is given in Chapter 6.

Chapter 2

Background

The following sections contain background information relevant to understanding the remainder of this thesis. A brief overview of melanoma's global impact is given, followed by a brief overview of skin anatomy relevant to understanding the different stages of melanoma. Current clinical and computer methods for detecting melanoma are given, with an emphasis on work related to this thesis. Image processing research tools used throughout this thesis are explained, with links to the sections in which they are used. The section concludes with a final note on the current trend in diagnostic smartphone apps.

2.1 Melanoma and the Skin

A brief overview of melanoma and how it relates to skin anatomy is presented in the following subsections.

2.1.1 Prevalence

Although melanoma causes 75% of all skin-cancer related deaths [1], only 5% of all diagnosed skin cancer cases are melanoma [4], and has many different surface-level appearances, many of which appear similar to benign lesions. These factors make it difficult to accurately identify malignant melanoma cases while maintaining a reasonable true negative rate.

Patient prognosis is highly dependent on the stage in which the melanoma is identified. Upon suspicion of a malignant lesion, the standard treatment involves physically excising

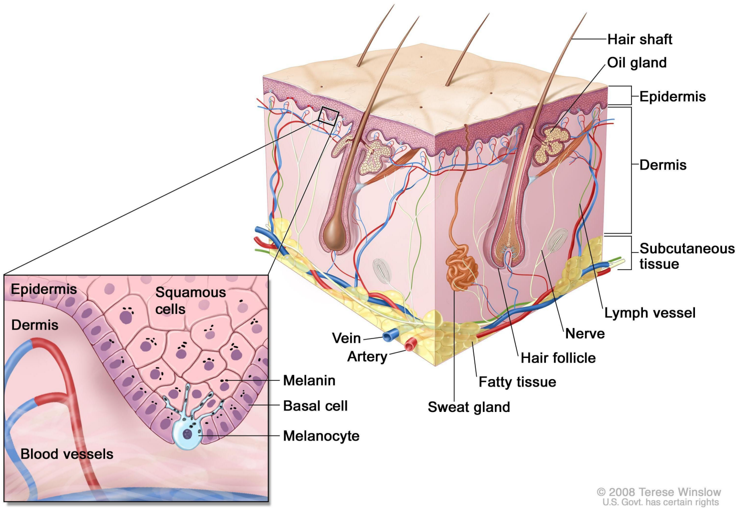


Figure 2.1: Simple anatomy of the skin[1]

the cancerous tissue. Reported five-year survival rates are 98% if melanoma is caught in situ (i.e., it has not spread), 62% if it has spread to regional organs, and a dismal 15% if it has spread to distant organs [4].

Although melanoma may metastasize areas other than the skin (e.g., eye, intestines), in the literature the term "melanoma" is usually indicative of *cutaneous melanoma* — that is, melanoma of the skin.

2.1.2 Skin Anatomy Overview

This section presents a brief overview of the anatomy of skin. The overview is constrained to identifying important high-level concepts pertaining to the onset and growth of melanoma. Detailed scientific skin anatomy and physiology is out of the scope of this thesis, though one can turn to a histology textbook for such information (e.g., [7]). Much of the material presented in this section is summarized from [7].

Skin Layers

Human skin is composed of two layers (see Figure 2.1): epidermis and dermis. Below the dermis lies the hypodermis, an integumentary layer which is important for understanding

[1] http://www.cancer.gov/cancertopics/pdq/treatment/melanoma/Patient/page1

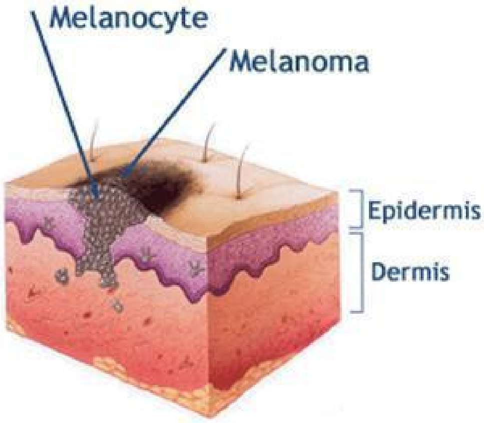


Figure 2.2: Visual depiction of melanoma growth[2]

melanoma.

The *epidermis* is the outermost layer of the skin. In many ways, it provides a protective barrier to and from the body. Melanocytes are situated at the bottom of the epidermis, and in the presence of ultraviolet (UV) radiation produce the pigment *melanin* (see below).

The *dermis* is much thicker and tougher than the epidermis in most areas of the body. Among its many contents are blood vessels, which circulate and replenish cells in the dermis and epidermis, and lymphatic vessels. It is through these vessels that melanoma can spread during the vertical growth phase (see Section 2.1.3).

The *hypodermis* (a.k.a. *subcutaneous tissue*, *subcutis*) acts as the interface between our skin and the rest of our body. It attaches the skin to bone and muscle, and provides a medium for larger blood and lymphatic vessels.

Melanin and Melanocytes

Melanocytes and melanin are important components to understanding melanoma. Melanocytes are cells located at the bottom of the epidermis (see Figure 2.1). In the presence of UV light, melanocytes produce melanin through a process called *melanogenesis*. Melanin is the pigment that determines a person's skin colour, including the effects of a "tan". Melanin has a large absorption band in the UV range of the electromagnetic (EM) spectrum [8], thus restricting the entry of damaging UV radiation into the hypodermis. UV

[2] http://www.nzmu.co.nz/What_Is_Melanoma_25.aspx

exposure has been linked to DNA photodamage, which can lead to cancerous replication of melanocytes, leading to melanoma (see Figure 2.2).

2.1.3 Growth Phases

The course of melanoma generally follows two growth phases [9]:

1. Radial growth phase (RGP)
2. Vertical growth phase (VGP)

The RGP usually precedes VGP. During the RGP, tumour cells spread outward in a radial fashion throughout the epidermis. At this point, melanocytes and melanin are still restricted to the epidermis (i.e., the cancer has not metastasized), so it is deemed “in situ”. For optimal prognosis, melanoma should be caught early during its radial growth phase, before the cancer enters a metastatic VGP.

During the VGP, melanocytes invade deeper into the dermis, and may spread to surrounding tissues via metastatic events. This spread to remote tissues is very dangerous, as the cancer can grow in many different parts of the body. As with most cancers, once it has spread to many areas, is it very difficult to localize and stop the transport of the cancerous cells. Lesions undergoing vertical growth usually appear on the surface as a bump, or nodule. This is especially common in nodular melanoma. As the VGP is the most dangerous one, such lesions should be addressed immediately.

2.1.4 Types of Melanoma

There are four primary types of cutaneous melanoma [9]: superficial spreading melanoma, nodular melanoma, lentigo maligna melanoma, and acral lentiginous melanoma. Each type of melanoma encompasses its own set of characteristics that are apparent from the lesion’s growth pattern. Refer to Table 2.1 for a synopsis of the various types of melanoma.

2.2 Diagnosing Melanoma Clinically

There are several phases leading up to definitive melanoma diagnosis. A brief overview regarding how melanoma is traditionally diagnosed in North America is given in the following subsections.

Table 2.1: Types of melanoma

Type of Melanoma	**Relative Frequency**	**Description**
Superficial Spreading Melanoma	70%	very irregular pigmentation; usually evolves from dysplastic nevus; characterized by radial growth
Nodular Melanoma	10–15%	rapid vertical growth; may appear where a lesion did not previously exist; most symmetric and uniform melanoma
Lentigo Maligna Melanoma	10–15%	often large; may have areas of hypopigmentation; typically found in sun-exposed areas
Acral Lentiginous Melanoma	5%	found in palms, soles, subungual areas; rapid progression from radial to vertical growth

2.2.1 Clinical Workflow

The typical way in which melanoma is identified and treated in North American is a multi-step process. Adults are encouraged to visit a general practitioner (GP) annually for a full physical analysis. During this time, the GP should visually check the entire body for any skin lesions that did not previously exist. This check can also be done by the patient on their own time, although an untrained person may neglect small subtle signs of growing melanoma, and may not be able to analyse their entire body effectively. If a suspicious lesion is identified, the GP will refer the patient to see a dermatologist, who specializes in matters of the skin. The dermatologist will usually visually analyse the skin lesion either with the naked eye, or with a dermatoscope (see Section 2.2.3). If the lesion is suspect of being melanoma, the dermatologist will excise around the lesion, and may send a biopsy to a pathologist, who can diagnose if it is indeed melanoma, as well as the specific type, using histopathological methods. In severe cases where the melanoma may have spread to remote tissues, more intense treatment may be administered, including chemotherapy and lymphadenectomy (removal of lymph nodes).

2.2.2 The ABCDs of Melanoma

Although several visual metrics exist for identifying melanoma, one of the most widely used ones is the ABCD metric [10, 11]. In this analysis, dermatologists attempt to identify:

- **A**symmetry (of colour and structure)
- **B**order irregularity
- **C**olour
- **D**iameter

The original weighting scheme [10, 11] is as follows. Note that utilisation methods are very subjective and can vary from doctor to doctor.

Asymmetry Asymmetry is determined with respect to both colour and structure patterns. Two orthogonal axes are determined by the dermatologist which produce the smallest amount of visual asymmetry. A score of 1 is assigned to each axis if it produced asymmetry, and 0 otherwise. The final score is calculated as the sum of the two axis scores. Thus, the final asymmetry score yields $A \in \{0, 1, 2\}$.

Border irregularity Irregularity with respect to shape and pigmentation is identified along the border. The lesion is visually split into eight radial segments, each of the same size, according to eight axes at 45° intervals. A score of 1 is assigned to each segment if it portrays border irregularity, and 0 otherwise. The final score is calculated as the sum of the axis scores. Thus, the final border irregularity score yields $B \in \{0, 1, 2, \ldots, 8\}$.

Colour Particular colours have been historically observed in melanomas. This colour distribution is easily identified using a dermatoscope (see Section 2.2.3). The number of unique colours in the lesion is identified and counted. Possible colours include white, red, light/dark brown, blue-gray, and black pigments. Thus, the final colour score yields $C \in \{1, 2, \ldots, 6\}$.

Diameter If the lesion is more than six millimetres in diameter, it is highly likely to be melanoma. This large radius is a by-product of the radial growth phase (see Section 2.1.3).

2.2.3 Equipment

The identification of cutaneous melanoma is a very visual process. Many dermatologists simply look at the area of interest on the skin and identify certain characteristics based on naked-eye observation. In fact, people are encouraged to routinely perform self-administered skin exams to spot any abnormal or new growths. However, some clinics use medical equipment to complement or augment the visual process. The most widely used equipment is briefly discussed here. A good review on existing technologies can be found in [12].

Dermoscopy [13] (also known as *epiluminescence microscopy* (ELM) or *dermatoscopy*) involves the use of a *dermatoscope*. A dermatoscope is a handheld optical tool used by some dermatologists. Although specific functions vary between models, most dermatoscopes are capable of magnification and, perhaps more importantly, discarding skin surface reflectance to elucidate sub-surface structures. This is usually done by applying dermoscopic oil to the skin, or by cross-polarization of the reflected light. Unfortunately, the clinical use of dermatoscopes is limited in North America, with one survey reporting less than 50% utilisation in the USA [3].

Multi-spectral imaging has seen some recent activity in skin cancer detection [14]. These imaging techniques detect the reflectance characteristics at different bands of the EM spectrum. This information can be used knowing the skin's EM absorption, transmittance, and reflectance spectra. By analysing the reflectance at multiple frequencies, non-visible information be inferred. For example, melanin, hemoglobin and collagen densities can be estimated [15, 16, 17]. Multi-spectral imaging devices can provide pertinent information that may not be directly visible. This advantage has led to the development of several commercial medical devices, such as Verisante Aura [18] and MelaFind [19].

Optical coherence tomography (OCT) [20] is based on interferometry, producing cross-sectional images (i.e., image "slices" of a certain thickness). OCT is widely used to examine retina (found in the eye) integrity, and feasibility studies are currently being performed for skin lesion diagnosis [21, 22]. Skin OCT devices can penetrate deep into the epidermis and the top of the dermis (see Section 2.1.2), elucidating vertical growth. However its resolution does not allow for analysis on the cellular level [14]. OCT is particularly useful for determining the depth of melanoma penetration, which is indicative of vertical growth (see Section 2.1.3).

2.3 Detecting Melanoma Automatically

Decision support systems for detecting melanoma generally follow the workflow outlined in Figure 2.3. Each stage is explained in the following subsections, along with a brief literature overview of existing methods relevant to this thesis. The contribution of this thesis relates to feature extraction, which is explained in Section 2.3.3.

2.3.1 Preprocessing

In skin cancer decision support systems, preprocessing involves illumination correction. Since the images are obtained in a natural setting, the illumination across the image is privy to factors such as natural lighting, office lighting, and camera flash. The goal of illumination correction is to adjust the original image data (i.e., pixel values) to standardize the lighting exposure across the entire image. This increases the reliability of the following steps, which depend on pixel values. Figure 2.4 presents a visual example of the effect of preprocessing using [23].

Illumination correction is important for the accuracy of features that rely on pixel values. Thus, preprocessing was a necessary step when validating the work embodied in this thesis. We used the illumination correction from [23] prior to feature extraction.

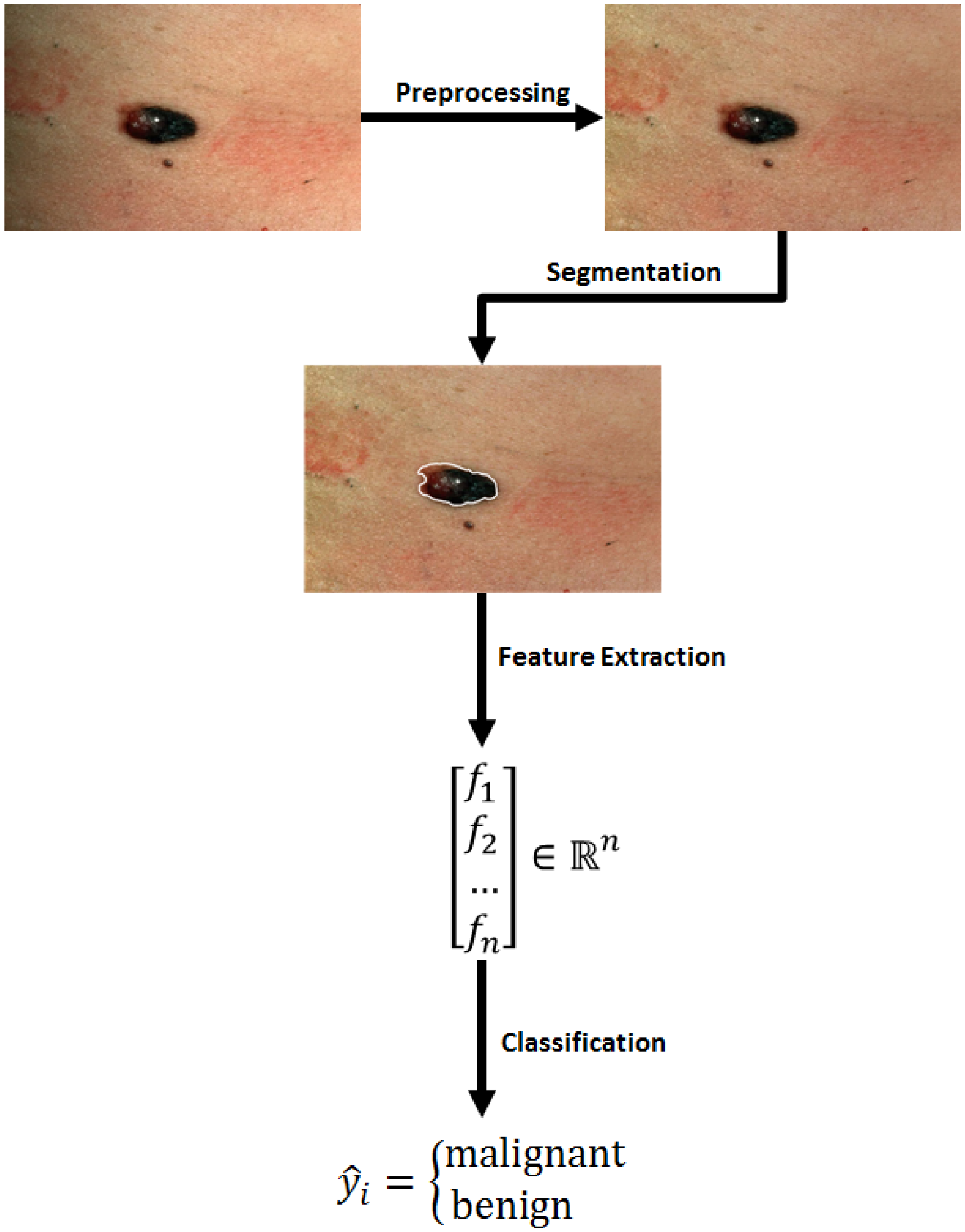
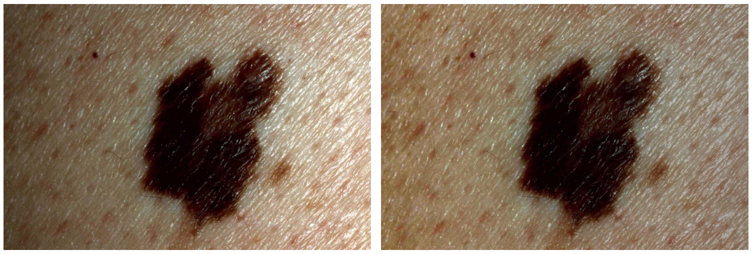

Preprocessing
Segmentation
Feature Extraction
$\begin{bmatrix} f_1 \\ f_2 \\ \cdots \\ f_n \end{bmatrix} \in \mathbb{R}^n$
Classification
$\hat{y}_i = \begin{cases} \text{malignant} \\ \text{benign} \end{cases}$

Figure 2.3: Classical skin cancer decision support system workflow. This thesis addresses the “feature extraction” step.



(a) Original image

(b) Image after preprocessing using [23]

Figure 2.4: Example of illumination correction during the preprocessing step. Notice how the preprocessed image is affected by non-uniform lighting. Preprocessing attempts to correct for this non-uniform lighting distribution using the existing pixel values.

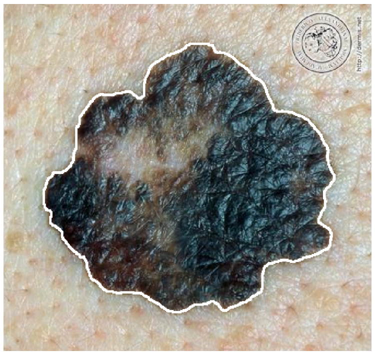


Figure 2.5: Example segmentation of a pigmented skin lesion. The segmentation separates the skin lesion from the surrounding healthy tissue.

2.3.2 Segmentation

Segmentation [24] is the process of characterising the image's pixels into semantic groups. In the case of skin lesion analysis, segmentation involves determining the border that separates the skin lesion from the surrounding healthy tissue. The result is a binary mask outlining the skin lesion. Figure 2.5 presents a visual example of ground-truth segmentation.

For skin lesion analysis, a segmentation algorithm will at best match a manually-defined segmentation. Our data set (discussed further in Section 5.1) comprises images that have been manually segmented and are used for feature extraction. Therefore, segmentation is not discussed further in this thesis.

2.3.3 Feature Extraction

Feature extraction [24] involves performing specific pre-defined calculations on the preprocessed and segmented image. The goal is to generate a feature vector (i.e., a vector of real numbers) from the image that aptly describes important characteristics of the image. Thus each image is represented by a point in some n-dimensional space (called the *feature space*). The goal of feature extraction is to construct a feature space such that the inherent image classes (e.g., malignant and benign) are easily separable in this space. Mathematically, given an image I, feature extraction is the mapping $f : I \mapsto F \in \mathbb{R}^n$, where F is the computed feature vector.

Many skin cancer feature extraction methods propose features that model the ABCD criteria used by doctors in clinical settings [25, 26] (see Section 2.2.2). Existing feature extraction methods are discussed in detail in Section 2.5.

Most methods have been designed for and validated against images obtained using a dermatoscope. As discussed in Section 2.2.3, dermatoscopes produce images with minimal skin surface reflection and elucidate sub-surface texture information, offering preferred images for image processing. However, since the features are designed for this specific application, they are not readily usable in a decision support system for analysing standard camera images. Cavalcanti and Scharcanski evaluated several state-of-the-art dermoscopic image processing techniques with their standard camera image dataset and found suboptimal results [27]. There are some papers that propose features for standard camera images [27, 28, 29], however the focus of these papers is either on the preprocessing and segmentation or classification phases. As a result, the feature sets are fairly basic, consisting of many low-level features that are combined to try to approximate ABCD. Work by Amelard *et al.* comprised within this thesis focused on feature extraction in this setting [30, 31, 32].

2.3.4 Classification

In machine learning and pattern recognition [33], the "classification" stage aims to assign a *label* (i.e., a *class*) to an image based on evaluating its feature vector. Skin cancer detection primarily involves *supervised learning*, in which a classifier is trained knowing the ground-truth labels of the training data. There are two phases to classification: *training* and *testing*. In the training phase, given training data in the form of feature vectors with ground-truth labels, the classification scheme attempts to learn a mapping for discerning classes from each other. In the testing phase, this mapping is used on new inputs to generate an estimate class label for the image. Mathematically, classification is the mapping $C : F \mapsto L$ where F is the image feature vector and L is the class label.

Most skin cancer detection systems have used existing “out-of-the-box” methods from machine learning. Common schemes in literature include support vector machines (SVM) [34, 35], artificial neural networks (ANN), decision trees, and k-nearest neighbour (K-NN) [33] (see [26] for a good literature overview). SVM is discussed further in Section 2.6.6, as it is used in the experimental setup discussed in Chapter 5.

2.4 Types of Images

The majority of skin cancer detection research has focussed on dermoscopic images (i.e., images obtained using a dermatoscope – see Section 2.2.3). However, recent surveys have indicated that dermatoscopes are not widely used in the USA for melanoma detection [3, 36]. The surveys do not indicate what percentage of dermatoscope users use digital dermatoscopes (i.e., dermatoscopes capable of capturing a digital image), however it can be postulated that given the dermatoscope market, far fewer digital dermatoscopes are being used than purely optical dermatoscopes. This poses a convincing argument for the restricted clinical use of dermoscopic image analysis.

We therefore turned to developing methods of evaluating images obtained using standard consumer-grade cameras. Some benefits of these types of devices are ease-of-use and low cost, leading to low barriers to adoption. However, these types of images pose great technical problems, as they are uncalibrated devices used in unconstrained environments. Some of the problems include variations in RGB colour representations, resolution, flash exposure, rotations, scale, and angle at which the image is captured. The features proposed in this thesis have been designed to mitigate the effects of these variations.

2.5 Related Work

Most feature extraction methods proposed in skin cancer detection literature have focused on designing features that model the ABCD criteria (see Section 2.2.2). Good reviews of existing features can be found in [26, 25].

Lee and Claridge proposed border irregularity indices [37]. Aribisala and Claridge proposed quantifying border irregularity from conditional entropy [38]. Celebi *et al.* proposed a set of features describing shape, colour, and texture with some rationale, recognizing the problem that feature sets were being presented without rationale [39]. This is an issue addressed in this thesis.

Most proposed colour features are, unfortunately, calculated in the RGB domain, which is not perceptually uniform (see Section 2.6.1). That is, the distance between two points in the three-dimensional RGB space is not indicative of their perceptual disparity. Furthermore, often authors will try to model (in RGB) colours commonly perceived in melanomas using a dermatoscope [28, 27]. However, cameras are not calibrated to a standard white point, and so these RGB values are not representative of calibrated colours. That is, a certain real-world colour may take on different values of RGB depending on the camera's sensor setup. Some methods use a perceptually-uniform colour space (e.g., CIE $L^*a^*b^*$), however most of the features themselves are very simplistic, such as measuring the minimum, maximum, average, and standard deviation of the colour channel values (see [26, 25] for overviews of existing features). There is a lack of robust colour features.

The methods discussed above have been designed to extract features from images obtained using a dermatoscope, which are generally not suitable for the highly-varying unconstrained conditions of standard camera images. To that end, there has been a limited amount of research dedicated towards proposing features for images obtained using standard camera images. The papers that have proposed feature sets for these types of images have focused mainly on the hard preprocessing and segmentation problems, resulting in low-level features. For example, Cavalcanti and Scharcanski [27] proposed the same set of low-level features as proposed by Alcon *et al.* [28] albeit some minor modifications. The first sets of high-level features were proposed by Amelard *et al.* for describing asymmetry and border irregularity [30, 31, 32]. These works are part of the main contribution of this thesis.

2.6 Image Processing Tools

This section provides an overview of image processing techniques used throughout this thesis. Each tool is linked to the section in which it is employed, for ease of comprehension.

2.6.1 Colour Spaces and Perceptual Uniformity

Colour spaces, such as the popular RGB (red-green-blue) space, are usually three- or four-dimensional spaces that maps a physical colour to a coordinate in the space, where each axis represents a certain characteristic for that colour space. Most images are expressed in the RGB colour space, where a colour is represented by "adding" certain amounts of red, green, and blue light to obtain a pixel's colour. Colour displays are generally made up of individual pixels, each of which are in turn comprised of three RGB light sources. It is therefore a natural fit to use the RGB colour space for image colour representation, as it is a direct mapping to the output's display.

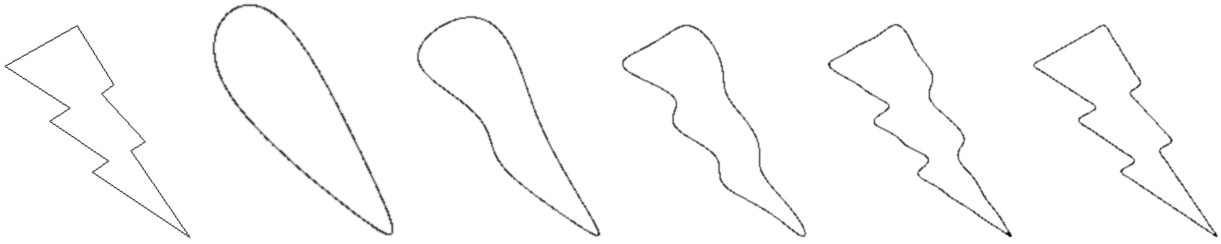
The CIE XYZ colour space [40] is a foundational one in the field of colour perception. The CIE XYZ space was modeled to fit the results obtained from colour perception experiments [41, 42]. The average human eye is able to observe colour using three different types of *cone cells*, which have overlapping sensitivity curves that peak in the blue, green, and red wavelength bands of light. The goal of the CIE XYZ colour space was to define a space that can represent every human-observable colour. Many colour models have been built upon the CIE XYZ space to take advantage of the perceptual realism.

A colour space is deemed "perceptually uniform" if the distance between any two points in the space (i.e., distinct colours) approximates the amount of human-perceived difference. Examples of such spaces include the CIE $L^*a^*b^*$ and Lch spaces, each derived from the CIE XYZ colour space. A perceptually uniform colour space is especially useful when comparing colours akin to human's perceptual difference, such as finding images that look similar to a base image. Section 4.1.1 uses CIE $L^*a^*b^*$ space to analyse the amount of perceptual colour asymmetry of a lesion. Section 4.3.2 uses the same colour space to analyse the amount of perceptual colour complexity in a skin lesion.

It is important to note that the RGB space is *not* perceptually uniform, and thus is not suitable for directly comparing colours (usually using the Euclidean distance). When dealing with such an application, CIE $L^*a^*b^*$ is often the best choice, which is a direct mapping from the RGB domain.

2.6.2 Fourier Descriptors

Fourier descriptors [24] are an application of Fourier theory onto shape analysis. The Fourier transform is widely used to transform a signal into the frequency domain, where the signal is represented as a weighted combination of sinusoidal frequencies. It is an invertible process, meaning that the frequencies can be uniquely mapped back to the original signal. Given a set of N complex numbers $\{x_i : x_i \in \mathbb{C}\}_i$, the discrete Fourier transform (DFT)



(a) Original (b) $n_f = 2$ (c) $n_f = 4$ (d) $n_f = 8$ (e) $n_f = 16$ (f) $n_f = 32$

Figure 2.6: Example of shape reconstruction using Fourier descriptors. The original border was sampled with 1000 points, and reconstructed using different numbers of frequencies (n_f). Each reconstruction uses the lowest n_f frequencies.

and its inverse are given below:

$$F_k = \sum_{n=0}^{N-1} f_n \exp(-2\pi ikn/N) \tag{2.1}$$

$$f_n = \frac{1}{N} \sum_{k=0}^{N-1} F_k \exp(2\pi ikn/N) \tag{2.2}$$

where F is the set of contributions from the uniformly-spaced frequencies. Thus F_0 is the DC component (i.e., offset), F_1 is the amplitude of the first low-frequency sinusoidal contributing to the original signal, and so on.

Fourier descriptors are used in two-dimensional shape analysis as follows. The trick is to represent each two-dimensional (x, y) coordinate pair as a one-dimensional complex number: $x+iy$. The standard DFT can be applied to this formulation to retrieve the shape's frequency decomposition. By omitting certain frequencies (i.e., setting their amplitude to 0), the shape can be reconstructed using a specified bandwidth of information. This can lead to more robust shape analysis methods. This is used in Section 4.1.2 and Section 4.2.2, which present ways of analysing structure asymmetry and border irregularity given the lesion's shape. Figure 2.6 shows a graphical example of shape reconstruction using Fourier descriptors.

Fourier descriptors are a global transformation. That is, changing a single frequency component affects the entire reconstructed shape. This may be unwanted functionality when analysing local structure changes in shape analysis. Morphological operations, discussed next, is a set of techniques that allows for such local analysis.

2.6.3 Morphological Operations

Morphological operations are techniques for analysing geometric shapes using set theory [43]. Each operation relies on the definition of a *structuring element*, which is usually some simple geometric shape, such as a disk or a line. In image processing, structuring elements are binary images modeling such a shape. They are used to modify the shape in some deterministic way.

Of particular importance to this thesis are *morphological opening* and *morphological closing*. Given a binary image A and a structuring element B, morphological opening and closing are defined as follows:

$$M_o(A,B) = \bigcup_{B_x \subseteq A} B_x \tag{2.3}$$

$$M_c(A,B) = M_o(A^*,B)^* \tag{2.4}$$

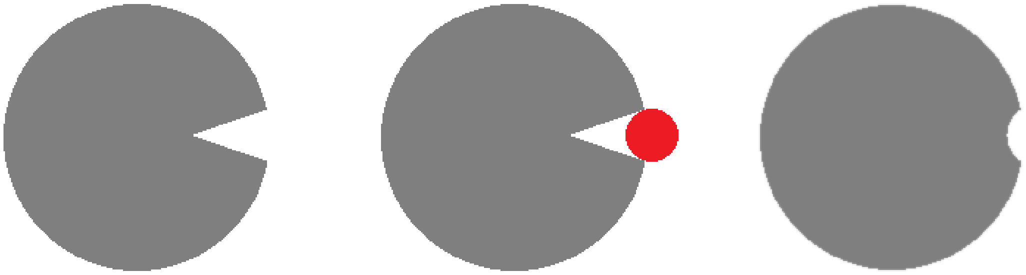
where X^* is the complement of X. An intuitive description of morphological opening and closing using a disk structuring element may be easier to comprehend. Given a shape A, if we "roll" the disk over the shape, the disk will roll over some areas without falling into them. The amount of area that it "rolls over" is dependent on the size of the shape and structuring element. If we fill in that area, we obtain the product of morphological closing. Similarly, if we roll the disk on the interior of the shape, and fill in any gaps in which the disk does not fall, we get the product of morphological opening. A visual example of this procedure is given in Figure 2.7.

Morphological operations, like Fourier descriptors, are very useful for analysing shapes in a robust and consistent manner. They are used in Section 4.2.1 to analyse border irregularity by filling in/cutting out spiky variations in the border.

2.6.4 Earth Mover's Distance

The Earth mover's distance (EMD) was an existing "transportation" linear optimization technique that was introduced into computer vision field by Rubner *et al.* [44]. The EMD is a method for comparing two distributions, or *signatures* of distributions. A signature of a distribution is defined as a set of clusters in which the number of elements (e.g., pixels) belonging to each cluster is tracked. In fact, a histogram is a special case of a signature. A signature can be computed by, for example, k-means clustering.

Given two signatures from two images, the EMD calculates the minimum amount of "work" required to transform one signature into the other. This is especially useful



(a) Original (b) Structuring element (c) Result

Figure 2.7: Visual depiction of morphological operations (closing). The disk cannot be self-contained in the crevice, so it is "closed" out from the shape.

when image pixel values are expressed in perceptually uniform spaces, as discussed in Section 2.6.1. Signatures need not contain the same number of clusters, since the distance between clusters in the space are all that's required for computing the EMD. Metaphorically, clusters from one signature can be thought of as piles of dirt in the given space, and clusters from the other signature can be thought of as holes that need to be filled. Each cluster (dirt/hole) is as "large" according to the number of points belonging to that cluster. The work involved is therefore the transportation of a certain amount of dirt across a certain distance.

Mathematically, the EMD seeks to optimize the following formulation:

$$EMD(P,Q) = \min_F \frac{\sum_{i,j} F_{ij} D_{ij}}{\sum_{i,j} F_i j} \tag{2.5}$$

subject to the following constraints:

$$F_{ij} \geq 0 \tag{2.6}$$

$$\sum_i F_{ij} = y_j \tag{2.7}$$

$$\sum_j F_{ij} \leq x_i \tag{2.8}$$

where P and Q are signatures; F_{ij} and D_{ij} are the flow (i.e., amount moved) and ground distance from source cluster i to target cluster j; y_j is the size of target cluster j; and x_i is the size of source cluster i. The following constraints are therefore imposed:

- Equation 2.6: allow only one-way shipping.
- Equation 2.7: forces target clusters to be “filled”.
- Equation 2.8: restricts the amount transferred from source cluster i to its size.

Rubner *et al.* concluded that colour comparison works better by representing a colour image using *signatures* rather than histograms [44]. Signatures are compact representations, such as the k centroids after performing k-means clustering on the colour vectors. This leads to simplified weighted representations that are less computationally expensive than per-pixel analysis and more robust than histogram-based comparisons [44]. We used a fast implementation of EMD in our calculations [45].

The EMD was used in Section 4.1.1 and Section 4.3.2 along with the perceptually uniform CIE $L^*a^*b^*$ space. In Section 4.1.1, the EMD was computed to determine the amount of colour difference on two sides of the lesion, providing an asymmetry score. In Section 4.3.2, the EMD was computed to determine the amount of colour complexity introduced when reconstructing the lesion’s colour distribution with varying numbers of colours, providing colour complexity scores.

2.6.5 PCA-Part for k-means initialization

PCA-Part is a deterministic method for computing initial cluster centroids for the k-means clustering algorithm [46]. The authors argue that using PCA projection is an effective way to initialize k-means since k-means attempts to minimize the within-cluster sum of squares, and PCA’s first principal component is the direction in which the largest amount of data variance is observed, which is similar to the k-means optimization scheme.

This initialization method is used in the clustering stages for colour analysis in Section 4.1.1 and Section 4.3.2. Since both the initialization and k-means itself are deterministic, this provides a deterministic clustering process. This is important to eliminate any source of randomness, which would lead to inconsistent feature values. As the decision support system is designed to be used in a clinical setting, reliable and consistent output is of utmost priority.

The algorithm is briefly described here. The data are split k times in a hierarchical manner by projecting them onto the first principal component of the PCA decomposition. From this projection, the data are segmented via thresholding according to Otsu’s method [47], which minimizes the combined within-class variance. Thus, for each iteration, one

```
function PCAPart($X$,$k$)
    $C_j^* \leftarrow X$                                        ▷ selected cluster
    while $k > 0$ do
        SplitCluster($C_j^*$)
        $C_j^* \leftarrow \arg\max_{C_j} \sum_{x_i \in C_j} \|x_i - \mu_j\|^2$
        $k \leftarrow k - 1$
    end while
    return $C$
end function

function SplitCluster($C_j$)
    Project $z_i \in C_j$ to the largest principal component axis $C_j$.
    Split $C_j$ into two sub-clusters $\{C_j^1, C_j^2\}$ according to Otsu's method.
end function
```

Figure 2.8: PCA-Part using Otsu's Method

cluster (initially the entire data) is split into two clusters. After k iterations, the centroids of each cluster are used as k-means initial centroids. The step-by-step algorithm is given in Figure 2.8.

2.6.6 Support Vector Machines (SVM)

Support vector machines (SVM) [35] is a classification scheme that is widely adopted in many machine learning settings due to its robustness and intuitive theory. The basic SVM setup is a linear classification problem. That is, it tries to find a $(d-1)$-dimensional hyperplane in a populated d-dimensional feature space such that a new data point can be classified based on a linear combination of select existing data. These data points, called *support vectors*, are those points that most fully represent the resulting hyperplane [35]. Being a simple and robust well-known classifier, we used linear SVM in our experiments, as discussed in Chapter 5.

Assume for simplicity that two classes are linear separable. There exists an infinite number of hyperplanes that can separate these data. SVM aims to find the hyperplane that maximally separates the data. This hyperplane is termed the *maximum-margin hyperplane*, as it tries to maximize the margin (i.e., the distance between the support vectors and the

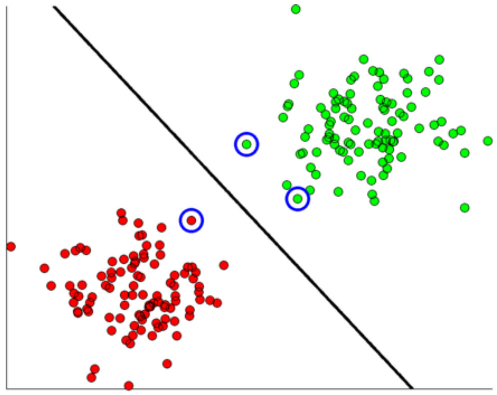


Figure 2.9: SVM maximum-margin hyperplane with linearly separable data. The support vectors are circled in blue.

hyperplane). Figure 2.9 shows a visual example of maximal-margin hyperplane in two-dimensional space.

Mathematically, given a set of labeled data points $\{(\mathbf{x}_i, y_i) : \mathbf{x}_i \in \mathbb{R}^d, y_i \in \{-1, 1\}\}_i$, the separating hyperplane can be formulated such that:

$$\mathbf{w} \cdot \mathbf{x}_s - b = \{1, -1\} \tag{2.9}$$

where $\mathbf{w}$ is the normal vector to the hyperplane, $\mathbf{x}_s$ is a support vector, and b is an offset. It can be shown mathematically that the distance between the two classes' support vectors is equal to $\frac{2}{||\mathbf{w}||}$, so to maximize the margin, we minimize $||\mathbf{w}||$. Recall that support vectors are the closest vectors from the two classes from which we want to maximize the margin. We therefore need to restrict any other data from being contained within the margin. Mathematically,

$$y_i(\mathbf{w} \cdot \mathbf{x}_i - b) \geq 1 \tag{2.10}$$

The trick here is that since $y_i \in \{-1, 1\}$, multiplying by y_i produces a positive number. Finding the maximum-margin hyperplane then becomes the following minimization problem:

$$\min_{\mathbf{w},b} \frac{1}{2}||\mathbf{w}||^2 \text{ s.t. } y_i(\mathbf{w} \cdot \mathbf{x}_i - b) \geq 1 \tag{2.11}$$

One more area we need to explain is the *soft margin*. It is seldom the case that real data is linearly (or even non-linearly) separable. That is, the amount of noise and variation in a system will inevitably lead to non-separability of the data. The "soft-margin" [35] extends on the original hard-margin SVM by modifying the original optimization formulation to incorporate what are termed *slack variables*. These slack variables allow for a certain amount of error for any given data point, effectively allowing the construction of a separating hyperplane containing a certain amount of misclassified data. Mathematically, Equation 2.10 becomes:

$$y_i(\mathbf{w} \cdot \mathbf{x}_i - b) \geq 1 - s_i \tag{2.12}$$

where s_i is a slack variable for data point $\mathbf{x}_i$. This leads to the following minimization problem:

$$\min_{w,s,b} \left\{ \frac{1}{2}||\mathbf{w}||^2 + C \sum_{i=1}^{n} s_i \right\} \tag{2.13}$$

subject to Equation 2.12, where C is a tunable parameter for controlling the impact of the slack variables on the optimization.

SVM can be made into a non-linear classifier by applying a projection kernel into the SVM formulation [34]. This kernel projects each point into a higher-dimensional space than the original feature space with the goal of being able to separate it in this synthetic space. This is a crucial field of study within SVM classification, but kernels are not used in this thesis as feature extraction should, at best, project the data such that they can be linearly separable. We therefore validate our work using linear SVM (i.e., emphasizing the performance of feature extraction rather than an advanced classifier). As it is out of the scope of this thesis, so it is not explored further.

2.7 Summary

The chapter has presented background material relevant to understanding the remainder of this thesis. The anatomy, physiology, and prevalence of melanoma have been briefly reviewed. Standard clinical methods for detecting and diagnosing melanoma have been reviewed. Relevant existing image processing research for automatic melanoma detection has been summarized, as well as existing image processing tools that will be used later in this thesis. In the next chapter, we present a framework for extracting intuitive features from melanoma images for automatic classification purposes.

2.8 A Note on Diagnostic Smartphone Apps

There has been a recent surge in melanoma detection smartphone applications ("apps"). However, the state of the literature for analysing standard camera images is still young.

Indeed, a recent study [48] evaluated the accuracy of four skin diagnostic apps. The authors of this study inputted digital clinical images of skin lesions with known histologic diagnoses into the various apps, and found highly variable accuracy metrics between the apps, with three of the four apps incorrectly classifying 30% or more of melanomas as benign. In the study, the best smartphone application was not an automatic detection process at all. Rather, it sends the image to an anonymous board-certified dermatologist, who provides an online diagnosis upon review of the obtained image. This study's results show that much work still must be done to increase reliability in automatic detection systems, and that we should not rely on apps that do not need regulatory approval.

Chapter 3

High-Level Intuitive Features

This chapter presents high-level intuitive features (HLIFs) as a framework for feature extraction for intuitive classification problems. The advantages of designing HLIFs are discussed, followed by general instructions for designing a HLIF. This framework is used in Chapter 4 for extracting features relevant to skin cancer detection.

3.1 Introduction

The success of a decision support system is highly reliant on the feature extraction stage. The goal of feature extraction in image processing is to calculate specific characteristics about an image such that the data is well-separated into its classes in the feature space. In the case of melanoma detection, the goal is to extract features from skin images such that malignant and benign cases are “easily” separated.

3.2 Problem: Low-Level Features

Low-level features are calculations used to describe a characteristic for which they were not designed. Many feature sets combine several low-level features to capture some high-level characteristic of an object. That is, one might use several ad-hoc calculations to describe a high-level characteristic. For example, to describe a lesion’s asymmetry, Cavalcanti and Scharcanski combine solidity, extent, circularity, equivalent diameter, and axis length ratios [27]. These features are standard shape analysis calculations that do not individually model lesion asymmetry. Combining them results in a high-dimensional feature space for describing asymmetry.

The benefits of using low-level features are that the features do not require significant design time, as they have already been conceived for other applications. However, the increased dimensionality of the feature space leads to many problems, such as curse of dimensionality [49], increased computational complexity, and possible overfitting due to the sparsity of the feature space. In fields that lack large amounts of data, this sparsity issue can be very problematic, as it is hard to show that the classifier is generalisable to new data. Furthermore, classification results using low-level features cannot easily convey intuitive rationale, as the features themselves are not intuitive to a human observer.

3.3 High-Level Intuitive Features (HLIFs)

We can now define a HLIF. A definition and reasons for following the HLIF framework for feature extraction are presented. Finally, instructions for designing a HLIF are given.

3.3.1 Definition

We define a *high-level intuitive feature* (HLIF) as follows:

High-Level Intuitive Feature (HLIF)

> A mathematical model that has been carefully designed to describe some human-observable characteristic, and whose score can be intuited in a natural way.

This is different than most features that are reused from previous applications. A HLIF captures a specific characteristic that is important to the given application (e.g., complexity of the colour distribution, smoothness of an object, etc.). This differs from most features in that HLIFs usually require more upfront design time, but they have the possibility of capturing relevant information about images that can be conveyed back to the user in an intuitive manner.

3.3.2 Why Use Them?

Many decision support systems aim to provide a classification for a new instance of an anticipated item. HLIFs are designed to model a human-observable characteristic, which provides two distinct advantages over common low-level feature sets:

1. HLIFs capture observable information.
2. HLIFs can be interpreted to provide the user with intuitive rationale.

Since HLIFs capture information that is relevant to the problem, usually fewer features are needed to accurately describe the data. This is explored in Chapter 5.

3.3.3 How to Design a HLIF

The first step to designing a HLIF is to study the target user with the goal of understanding how they analyse the data. Recall that HLIFs are modeled according to a *human-observable characteristic*. These characteristics are unique to each application. Depending on the circumstances, this can be done by directly studying the daily activities of a target user, or by conducting a literature review on methods used in practice.

Once a working understanding of the standard analysis techniques is complete, a review of fundamental mathematical methods should be conducted to determine which tools are available for modeling certain high-level characteristics. For example, if one of the characteristics involves describing the colour distribution of an object, one might turn to perceptually uniform colour spaces based on the CIE XYZ colour space, such as the CIE $L^*a^*b^*$ colour space.

Finally, once a thorough working understanding of existing mathematical methods is obtained, much deliberate thought must be put into the modeling stage. That is, the mathematical model of the feature must use existing tools in a way that characterizes a high-level characteristic, and in which the final result of the feature calculation can be conveyed back to the user in an intuitive manner (e.g., graphically).

Although this thesis focuses on the application of HLIFs for skin lesion detection, the HLIF framework can be applied to any problem that involves classification of data into intuitive classes.

3.4 Summary

This chapter has presented a framework for designing HLIFs for modeling human-observable characteristics in an image classification setting. In the next chapter, this framework is used to generate 10 HLIFs for skin lesion analysis.

Chapter 4

HLIFs for Characterizing Melanoma

This chapter presents the design and calculation of 10 HLIFs for the detection of melanoma in images obtained using standard consumer-grade cameras. These features were designed to model the ABCD metric widely used by dermatologists. A synopsis of the ABCD metric was given in Section 2.2.2. As the feature models follow the HLIF framework, the images can be queried by the user for intuitive diagnostic rationale. Note that the diameter ("D") was not modeled due to the unconstrained environment inherent in the data, making it very hard to infer physical scale in images. The experimental results using these features are discussed in Chapter 5.

4.1 Asymmetry

Dermatologists try to identify asymmetry of the shape and/or colour of a skin lesion. Benign nevi (i.e., "moles") are usually observed to have more even colour distributions and are more elliptical than malignant melanomas. Melanoma cases tend to have complex shapes with asymmetric colour distributions. These asymmetry HLIFs are extensions of work originally presented in [30, 32].

4.1.1 HLIF for Colour Asymmetry

A successful quantitative feature for describing colour asymmetry will be able to differentiate lesions based on the spatial uniformity and symmetry of the colour distribution.

Given a segmented skin lesion, an initial axis of separation (AoS) was set to the major axis, which passes through the centre of mass (i.e., centroid) of the lesion shape and describes the maximum amount of structural variation (i.e., the transverse diameter of the fitted ellipse). The colour distributions in the perceptually-uniform CIE $L^*a^*b^*$ space on each side of this AoS were compared. In particular, k signatures on both sides of the AoS were determined using k-means clustering, using the final k clusters as colour signatures. Mathematically,

$$S_i^\theta = k\text{-means}(C_i^\theta, k) \tag{4.1}$$

where θ denotes the orientation of the AoS, $S_i^\theta \in \{S_1^\theta, S_2^\theta\}$ is the colour signature (weighted clusters) in CIE $L^*a^*b^*$ space to either side of the AoS, $C_i \in \{C_1^\theta, C_2^\theta\}$ is the colour distribution to either side of the AoS in CIE $L^*a^*b^*$ space, and k-means(C_i^θ, k) is k-means clustering of data C_i^θ into k clusters.

Intuitively, S_i^θ is a set of points in three-dimensional space with masses equivalent to the number of points within the cluster. The Earth Mover's Distance (EMD) was computed using these two signatures (see Section 2.6.4). This "distance" metric quantifies the amount of perceptual work needed to transform the colour distribution from one side of the lesion to the colour of the other, thus effectively representing the amount of colour asymmetry. This formulation was repeated over n equally-spaced orientations so that a uniform sampling of AoS was considered. The feature calculation was determined to be the maximum asymmetry score yielded over the n trials.

Note that given a fixed initialization, k-means clustering produces deterministic clusters. In order to ensure consistent feature calculations, deterministic k-means initialization using PCA-Part with Otsu's method was performed [46, 47] (see Section 2.6.5). Consistency is important for doctor-system trust.

The final feature calculation is as follows:

$$f_1^A = \max_\theta \left\{ EMD(S_1^\theta, S_2^\theta) \right\} \tag{4.2}$$

where θ denotes the orientation of the AoS, S_1^θ and S_2^θ are the colour signatures in CIE $L^*a^*b^*$ space as in Equation 4.1. In our tests, we used $k = 10$ colour clusters and $n = 12$ separation axes.

Figure 4.1 depicts an example of this HLIF. The maximal AoS is plotted as a white line through the centroid of the lesion, and the obtained CIE $L^*a^*b^*$ colour signatures of both sides of the AoS are plotted, where the size of the sphere denotes the number of pixels belonging to that cluster centroid (i.e., weight). Figure 4.1b and Figure 4.1c intuitively capture the primary observable colours above and below the AoS. For example, Figure 4.1b shows primary dominance of light-tan colours as well as smaller concentrations of dark-brown colours. It can be observed that there is a significant amount of work required to transform the spheres in Figure 4.1b to Figure 4.1c, which is captured using the EMD.

4.1.2 HLIF for Structure Asymmetry

As a lesion's shape deviates from the ideal elliptical structure, it becomes less likely that the shape is symmetric. That is, structural asymmetry can be approximated by the coarse complexity of the lesion's spatial structure. Using Fourier descriptors (see Section 2.6.2), we reconstructed the lesion shape in two low-frequency ways to quantify structure complexity, according to the following algorithm.

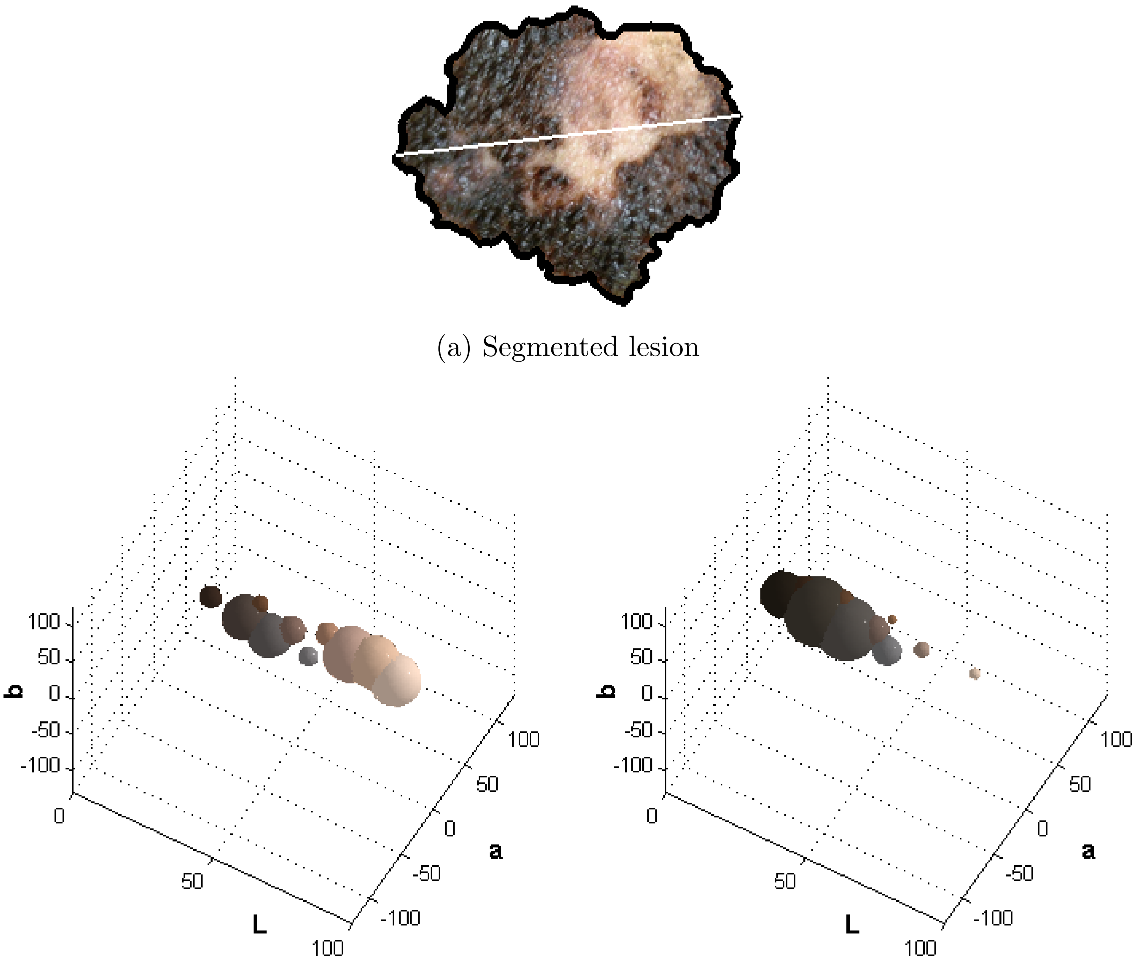
The lesion border was sampled using a pre-determined sampling rate. This is an important step, since the number of frequencies represented by the Discrete Fourier Transform is equal to the number of discrete spatial samples (see Equation 2.1). Using a constant sampling rate ensures consistent reconstruction. The Fourier descriptors of the shape were computed. Frequency components were discarded (i.e., their amplitudes were set to 0) in order to omit certain frequency information. The inverse FFT (IFFT) was used to generate two low-frequency reconstructions. The first reconstruction used the lowest two frequencies, which represents the most general approximation of the lesion border assuming a circular shape. The second reconstruction used several more frequencies to capture the presence of coarse structural variability, thus accounting for the structure's complexity, if any exists. The exact number of frequencies is dependent on the sampling rate. The normalized area between these two reconstructions was used to quantify the amount of complexity. Complex structures will exhibit large area differentials, and simple structures (e.g., oval-shaped benign lesions) will exhibit very little difference.

The final feature calculation is as follows:

$$f_2^A = \frac{area(R_{low} \oplus R_2)}{area(R_{low} \cup R_2)} \tag{4.3}$$

where $area()$ is a function that calculates the area of a shape, R_{low} and R_2 are the low- and two-frequency reconstructions, and $\oplus, \cup$ are the XOR and UNION operators. In our tests we used a 1000-point sampling rate and empirically chose five low frequency components.

Figure 4.2 depicts an example of this HLIF. It can be observed that the structure is very asymmetric, with one side containing a much smaller abnormal pigmentation density than the other. The coarse structure variation is captured in Figure 4.2 by the five-frequency reconstruction (green) and not the two-frequency reconstruction (pink). Thus, there exists a significant area differential between the two, indicating likely asymmetry.

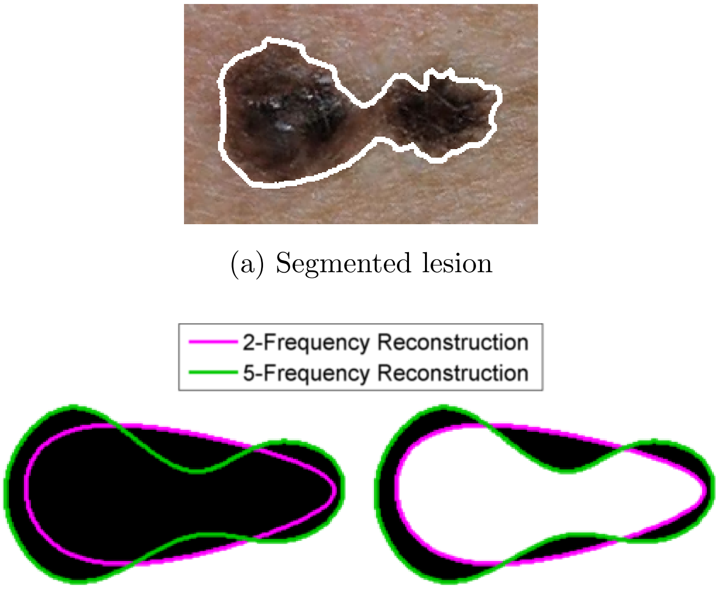


(a) Segmented lesion

(b) $L^*a^*b^*$ colour signature above the line

(c) $L^*a^*b^*$ colour signature below the line

Figure 4.1: Example of f_1^A on a superficial spreading melanoma with asymmetric colours. Notice how (b) and (c) capture the intuitive colour characteristics of the lesion on each side of the line, irrespective of texture and lighting variation. It is apparent that “work” is required to transform one colour signature into the other. In this case, $f_1^A = 23.86$.



(a) Segmented lesion

(b) Union

(c) XOR

Figure 4.2: Example of f_2^A on a superficial spreading melanoma with asymmetric structure. The asymmetry is introduced due to the lack of pigmentation density in the middle of the lesion. This structural variation is captured in the area differential between the two-frequency and five-frequency border reconstructions. In this case, $f_2^A = 0.327$.

4.2 Border Irregularity

Dermatologists try to identify irregular borders of the skin lesion. Benign nevi usually have very smooth elliptical shapes. Melanoma cases tend to deviate from this in terms of border pigmentation variations as well as “spiky” borders. These border irregularity HLIFs are extensions of work originally presented in [31, 32].

4.2.1 HLIF for Fine Irregularities

Melanoma cases often contain fine localized border irregularities in the form of abrupt localized pigmentation patterns, such as “spikes”. In order to quantify this information, we can use the existing theory of morphological operations (see Section 2.6.3).

Recall from Chapter 2 that morphological operations, unlike Fourier descriptors, are able to manipulate shapes on a *local* scale. The amount of localized abrupt pigmentation can be measured using morphological opening and closing. We compared the normalized difference in area resulting from these operations as compared to the original lesion. This captures abrupt (“fine”) irregularities in the border.

The final feature calculation is as follows:

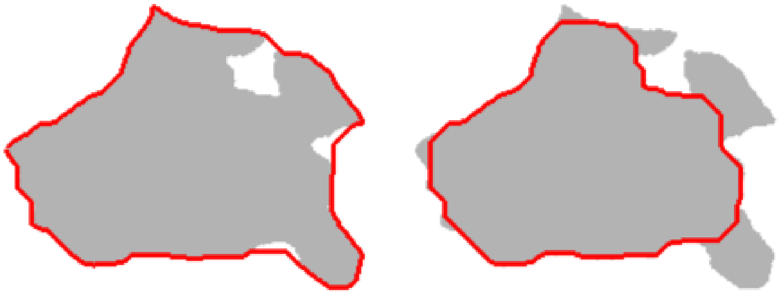
$$f_1^B = \frac{A_{closed} - A_{lesion}}{A_{lesion}} + \frac{A_{lesion} - A_{opened}}{A_{lesion}} \tag{4.4}$$

where A_{closed} and A_{opened} are the areas resulting from performing morphological closing and opening on the original lesion, and A_{lesion} is the area of the original segmentation. The two summed terms represent the amount of exterior and interior irregularities. In our tests, we used a disk structuring element of radius 20.

Figure 4.3 depicts an example of this HLIF. In Figure 4.3a, irregular valleys in the border are filled in, accounting for a significant proportion of the overall area. Similarly in Figure 4.3b, the irregular peaks are filled out. The amount of area filled in/out is a good indicator of border irregularities.

4.2.2 HLIF for Coarse Irregularities

Coarse border irregularities may also be present in melanoma cases. These irregularities are general structural shapes that deviate in a large way from an oval shape, such as large “swoops” in the border.



(a) Morphological closing (b) Morphological opening

Figure 4.3: Example of f_1^B on a lesion shape whose border contains fine irregularities (peaks and valleys). Morphological closing successfully fills in the abrupt valleys, and morphological opening fills out the abrupt peaks. In this case, $f_1^B = 0.208$.

From a signal processing perspective, these irregularities can be conceptualized as large deviations in low frequency information. It therefore seems natural to once more use Fourier descriptors to capture this information. In particular, much like the structural asymmetry HLIF presented in Section 4.1.2, the lesion was sampled at a pre-determined sampling rate and reconstructed using only a small amount of low frequencies. To quantify the coarse structural deviations, the perimeters of the low-frequency reconstruction and the original border lesion were compared.

The final feature calculation is as follows:

$$f_2^B = \frac{|P_{orig} - P_{low}|}{P_{orig}} \tag{4.5}$$

where P_{orig} and P_{low} are the perimeter lengths of the original and low-frequency reconstruction. We used 1000-point sampling rate and empirically chose four-frequency reconstructions.

Figure 4.4 depicts an example of this HLIF. Notice how the border has several parts at which it swoops down below and back up above the low-frequency border reconstruction. These are characteristic patterns of coarse border irregularities, where the border does not follow a smooth oval shape.

4.3 Colour Variations

Dermatologists try to identify specific colour patterns that have been historically found in melanoma cases. Unfortunately, most of the ABCD colour characteristics [10, 11] are

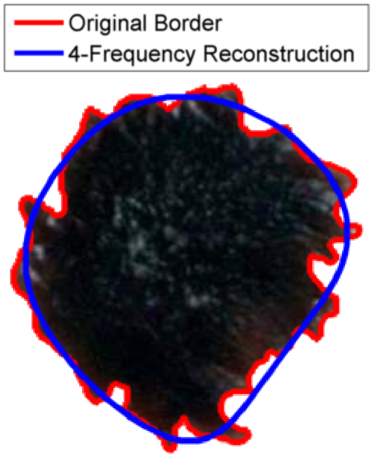


Figure 4.4: Example of f_2^B on a superficial spreading melanoma with coarse border irregularities. Notice the general deviations away from the smooth oval shape produced by the four-frequency reconstruction. In this case, $f_2^B = 0.325$.

only observable with the aid of a dermatoscope. Furthermore, colour medical imaging has not received the same amount of attention compared to monochrome medical image processing [50]. There is therefore a significant demand for novel research on quantifying colour information pertaining to melanoma detection, particularly using standard camera images.

4.3.1 A Colour Complexity Analysis Framework

As is apparent from the clinical definition, malignant melanomas tend to exhibit more complex non-uniform colour distributions than benign nevi. This can be attributed to the underlying metastatic growth of melanocytes, which in turn produces varying amounts of pigment. Exact colour patterns vary widely between cases, however the fundamental "complex" nature of the colour distribution can be observed in many melanoma cases.

The goal of these HLIFs is therefore to capture the *complexity* of the colour distribution. An intuitive way to determine this information is to compare and contrast reconstructions of the lesion's colour distribution using fixed numbers of representative colours. The presented theory for this framework draws from *sparse dictionary learning* research methods.

Consider the following cases as illustrative examples. A typical benign nevus has a fairly uniform colour distribution. It therefore stands to reason that the lesion's colour can be estimated fairly accurately by using the single most representative colour for a given lesion (i.e., a representative red colour). However, consider a case of melanoma that exhibits varying colours (see, for example, Figure 4.6a). It is impossible to find a single colour that accurately represents the lesion's colour distribution.

The colour complexity analysis framework comprises the following four stages:

1. Transformation of the image to a perceptually uniform colour space.
2. Construction of feature representations that model the colour information for a patch (i.e., local grid) of pixels.
3. Clustering the feature vectors into k colour clusters.
4. Quantifying the variance found using the original lesion and the k representative colours.

Step 1: Perceptual Uniformity

We transformed the original RGB image into the CIE $L^*a^*b^*$ space [40], in which the colour distribution is perceptually uniform (see Section 2.6.1). This should mitigate the effect of uncalibrated cameras, as the perceptual difference should be similar regardless of slight tonal differences between cameras. This way, we can compare colour values in a way that it mimics the amount of perceptual difference.

Step 2: Patch Representation

The goal of this step was to represent each patch of pixels in such a way that patches with similar pigmentation get grouped together in **Step 3**. To do this, two types of information were extracted from each patch: colour information and spatial information. This way, spatial constraints enforce locally cohesive colour structures, modeling the spatial localization of skin blotches. Intuitively, this can be represented by concatenating each column of pixel values in a patch across each CIE $L^*a^*b^*$ channel into a one-dimensional vector, and encoding the centre pixel coordinate for spatial context. Mathematically, for a given pixel patch $P_w(\mathbf{x})$ of width w centered around the pixel at location $\mathbf{x} = (\mathbf{x}_x, \mathbf{x}_y)$ in image I in CIE $L^*a^*b^*$ space, the spatial (f^s) and colour (f^c) feature vectors for patch P_w

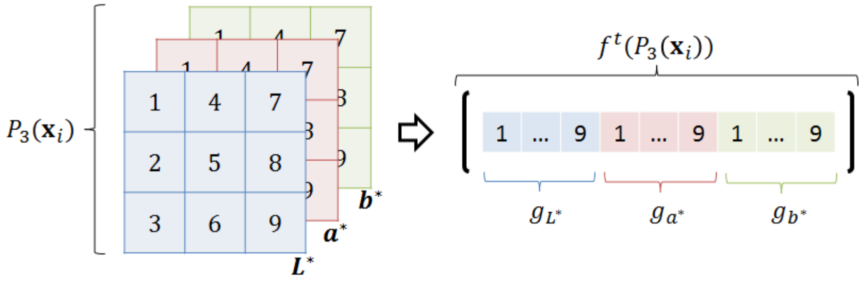


Figure 4.5: Graphical representation of the patch colour representation in Equation 4.7. For each channel, the pixel values are concatenated consecutively from 1 until 9.

were defined as:

$$f^s(P_w(\mathbf{x})) = \begin{bmatrix}\mathbf{x}_x & \mathbf{x}_y\end{bmatrix}, \tag{4.6}$$

$$f^c(P_w(\mathbf{x})) = \begin{bmatrix}g_{L^*}(P_w(\mathbf{x})) & g_{a^*}(P_w(\mathbf{x})) & g_{b^*}(P_w(\mathbf{x}))\end{bmatrix}, \text{where} \tag{4.7}$$

$$g_c(P_w(\mathbf{x})) = \begin{bmatrix}P_w(\mathbf{x}_{11}, c) & P_w(\mathbf{x}_{21}, c) & \dots & P_w(\mathbf{x}_{w1}, c) & P_w(\mathbf{x}_{12}, c) & \dots & P_w(\mathbf{x}_{ww}, c)\end{bmatrix} \tag{4.8}$$

where $P_w(\mathbf{x}_{ij}, c)$ is the pixel value from the channel c at the i^{th} row and j^{th} column in the patch. Note that $card(g_c(P_w(\mathbf{x}))) = w^2$, making $card(f^c(P_w(\mathbf{x}))) = 3w^2$ and $card(f^s(\mathbf{x_i})) = 2$, where $card(\cdot)$ is the cardinality function. A graphical depiction of f^c is given in Figure 4.5. The final colour feature vector was the concatenation of the spatial and colour information:

$$f^*(P_w(\mathbf{x})) = \begin{bmatrix}f^c(P_w(\mathbf{x})) & f^s(P_w(\mathbf{x}))\end{bmatrix} \tag{4.9}$$

Step 3: Finding Representative Colours

Upon populating the feature space with vectors $f^*(P_w(\mathbf{x}_i))$ for each point $\mathbf{x_i}$ in the image, k-means clustering was used to determine the k most representative colours of the lesion. Recall that k-means performs clustering by minimizing the within-cluster sum-of-squares distance of the clusters. This translates to clustering according to perceptual similarity. For consistency and reproducibility, PCA-Part using Otsu's method was used to generate a deterministic cluster initialization (see Section 2.6.5). Since the effect of the spatial characteristics in Equation 4.6 is affected by the patch size (i.e., the length of the feature vector), an additional weighting term was added into the k-means within-class sum-of-squares criterion as follows:

$$D = \arg\min_{S} \sum_{j=1}^{k} \sum_{f^*(\cdot)\in S_j} \Big(||f^c(P_w(\mathbf{x_i})) - \mathbf{d}_j^c||^2 + ||\lambda \cdot f^s(P_w(\mathbf{x_i})) - \mathbf{d}_j^s||^2 \Big) \tag{4.10}$$

where $D = \{\mathbf{d}_i\}_i$ is a set of k centroid colour-spatial elements, d^c and d^s are the colour and spatial components of the colour-spatial element, and λ is a relative spatial weighting term. The set $\{\mathbf{d}_i^c\}_i$ can be regarded as a set of representative colour patches.

Step 4: Colour Reconstruction

The output of **Step 3** is a set of k colours (i.e., the centroids of the clusters) along with each cluster's comprising pixels. Using this information, the image (lesion) can be reconstructed by replacing each pixel's original value with that of the representative centroid. These reconstructions can be used to quantify the amount of colour variation.

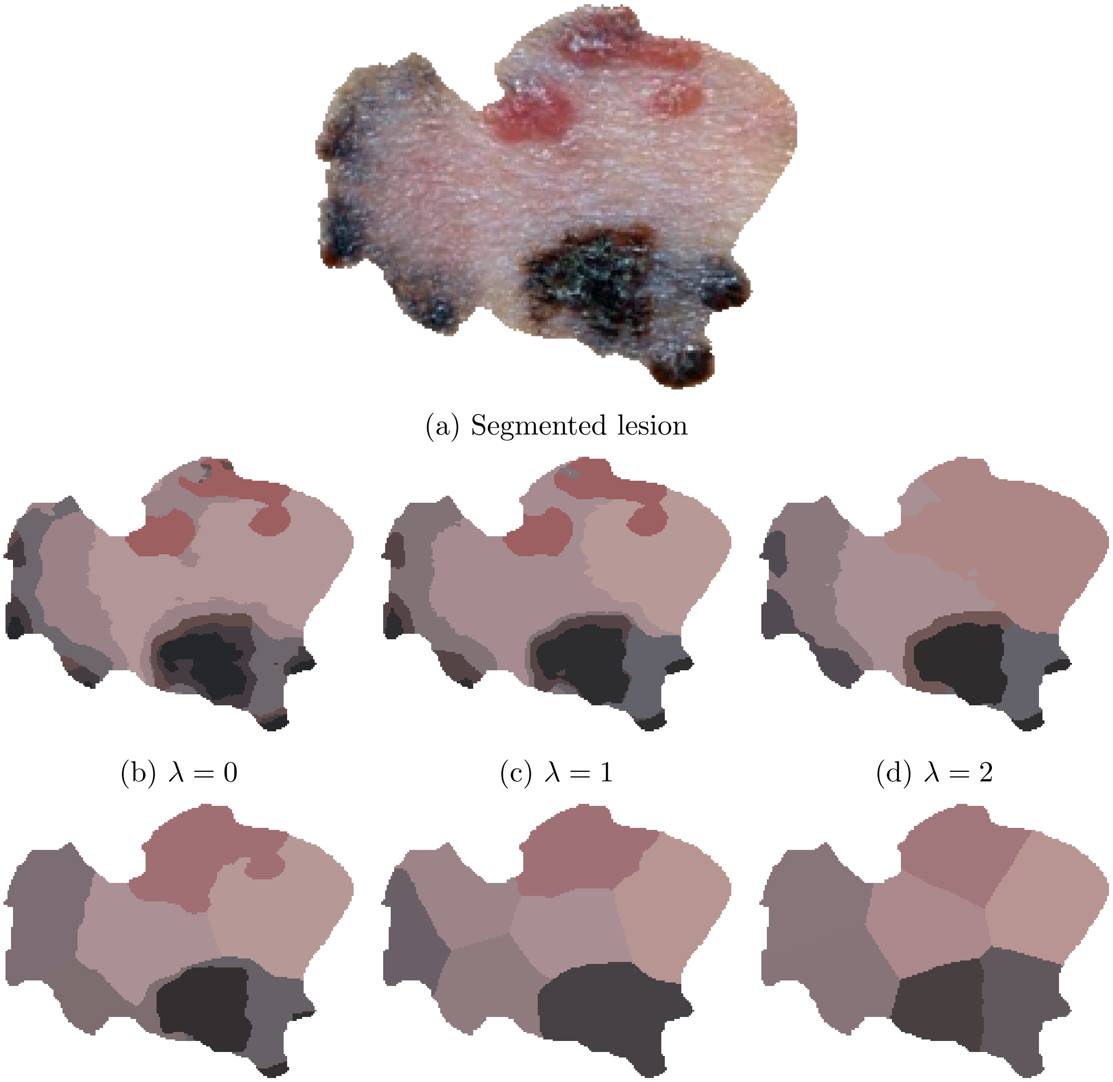
Effect of Parameters

This framework is influenced by three parameters: $\{\lambda, k, w\}$. The effects of each on the framework output are discussed below.

Effect of λ Figure 4.6 shows the effect varying the spatial weighting term λ on the colour reconstruction formulation. For illustration, we used seven clusters and 9×9 patches.

When there is no spatial weighting term (i.e., $\lambda = 0$), the clustering is solely based on the pixel colours (i.e., their CIE $L^*a^*b^*$ values). As seen in Figure 4.6b, this introduced very small localized colour attributes. As we are trying to model the overall colour distribution, this is not suitable for our needs.

On the other extreme with a large spatial weighting term (i.e., $\lambda = 16$), the clustering is heavily weighted on the distances between pixel locations. As seen in Figure 4.6g, the resulting reconstruction was largely segregated into approximate equal-sized blocks whose colour was determined from the underlying pixels. This is also not suitable for our needs, since it does not accurately portray the variations in colour across the lesion.



(a) Segmented lesion

(b) $\lambda = 0$

(c) $\lambda = 1$

(d) $\lambda = 2$

(e) $\lambda = 4$

(f) $\lambda = 8$

(g) $\lambda = 16$

Figure 4.6: Illustrative example of the effect of varying the spatial weighting term λ on a lesion with a complex colour distribution. As λ increased, we observed fewer (unwanted) small localized regions, and more self-contained regions representing colours within its spatial vicinity. Very high λ resulted in unrepresentative equally-sized regions. In this example we used seven clusters and 9×9 patch sizes.

Effect of k Figure 4.7 shows the effect varying the number of representative colour clusters (i.e., k in k-means clustering) on the colour reconstruction formulation. For illustration, we used $\lambda = 0$ and 9×9 patches.

When there is only a single cluster, the reconstruction is a single representative colour. In Figure 4.7b we can visually verify that the red-brown colour is a good general fit to approximating the original lesion colour. In the case of a lesion with uniform colour distribution, this colour will reconstruct the original colour distribution well.

On the other extreme with many colour clusters (e.g., 50 clusters), the reconstruction overfit the complex colours introduced by skin reflections (e.g., flash, lighting conditions, etc.). This is seen in Figure 4.7f, especially in the larger dark region. The reconstruction shows many light regions within the dark region, which are errors from overfitting to the observed skin reflectance colours.

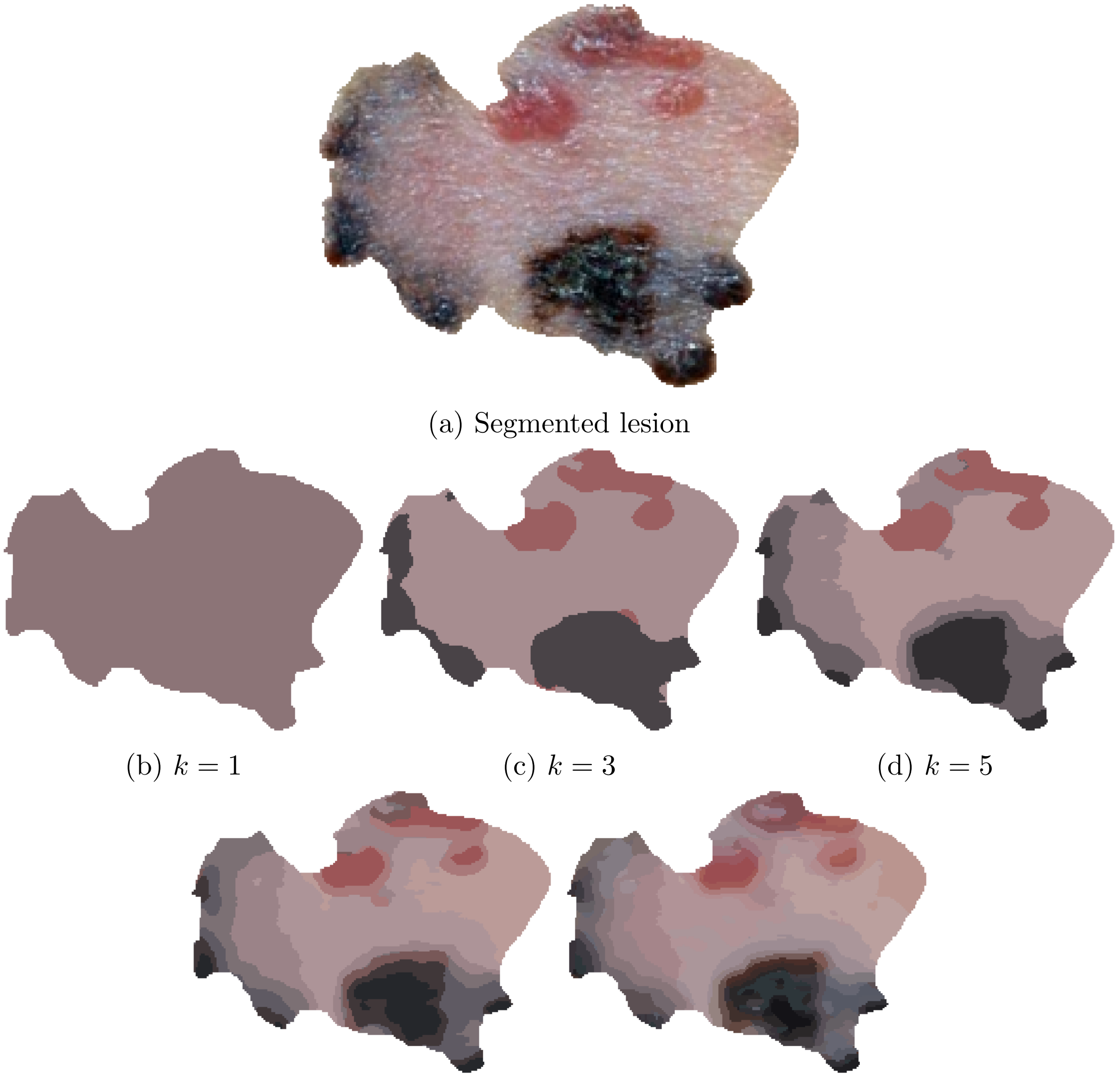
Effect of w Figure 4.8 shows the effect varying the patch width on the colour reconstruction formulation. For illustration, we used $\lambda = 0$ and seven clusters.

When the patch incorporates a single pixel (i.e., 1×1), no information about the surrounding pixels is considered, thus the original image is reconstructed pretty accurately. This behaviour is apparent in Figure 4.8b, which closely resembles the original colour distribution albeit being reconstructed with only seven clusters. This is not the behaviour that we want, as we are trying to reconstruct the underlying colour distribution independent of skin surface reflections.

On the other extreme with a large patch size (e.g., 17×17), patches incorporate information from within a large neighbourhood. Thus its variance may be too large to be useful. This is apparent in Figure 4.8f where colour regions seem to blend into others unrepresentative of the original colour space.

4.3.2 HLIFs for Colour Complexity

Remember, the goal is to use the reconstruction scheme from Section 4.3.1 to generate features whose models can be intuited. As aforementioned, “simple” benign lesions may be reconstructed accurately using as little as one representative colour, whereas melanoma colour distributions are more complex. We generated three sets of HLIFs to satisfy these conditions. In our tests, we empirically chose 9×9 patches, spatial weight $\lambda = 1.75$, and $\{1, 2, 5\}$ clusters. Although automatic parameter selection would be ideal, due to the time constraints of this thesis, these parameter values were chosen using anecdotal visual



(a) Segmented lesion

(b) $k = 1$

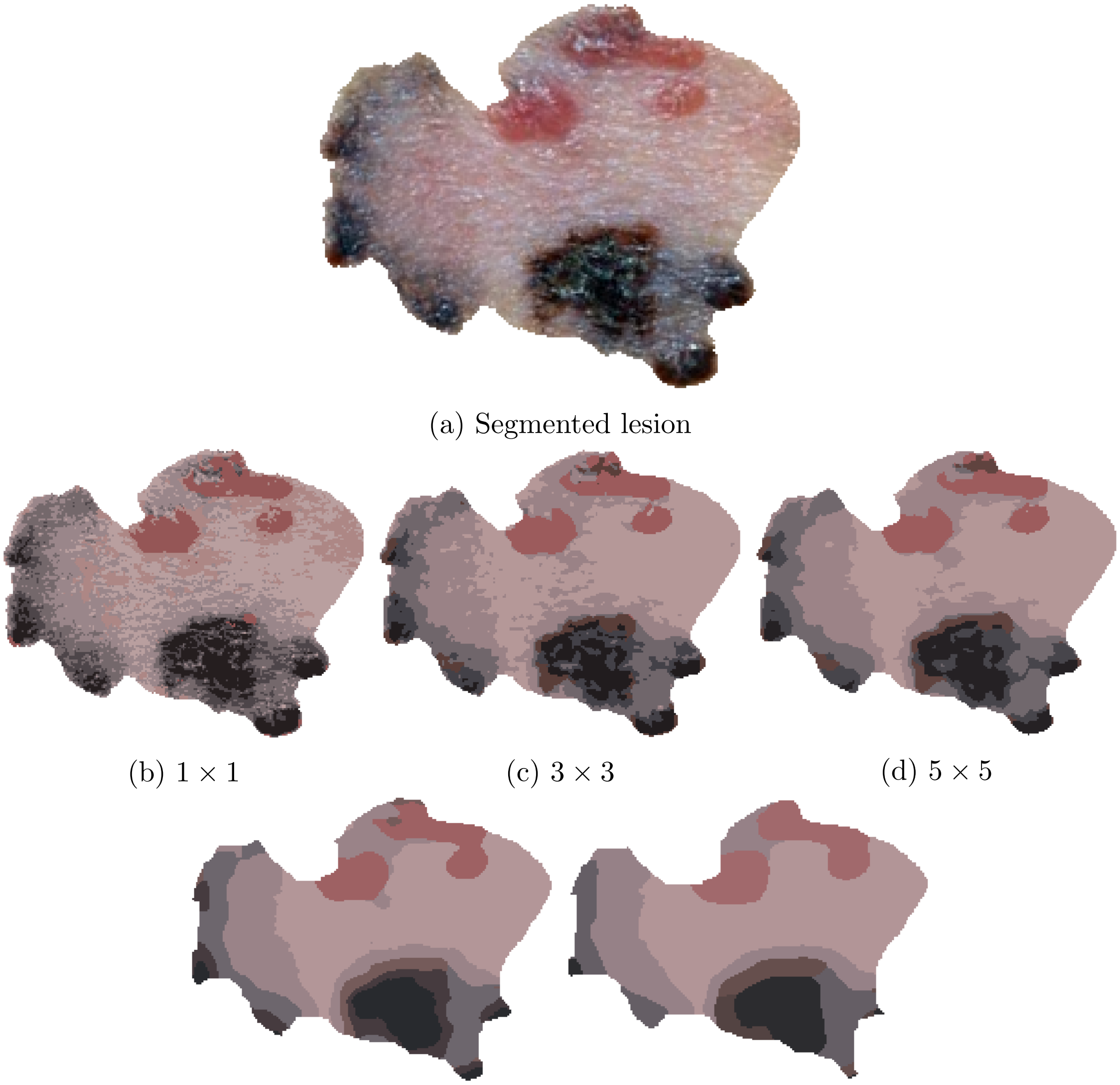
(c) $k = 3$

(d) $k = 5$

(e) $k = 10$

(f) $k = 50$

Figure 4.7: Illustrative example of the effect of varying the number of colour clusters used for reconstruction on a lesion with a complex colour distribution. Using a single cluster resulted in an overall representative colour. As the number of clusters increased, more colours from the original lesion were captured. Too many clusters resulted in overfitting skin reflectance. In this example we used $\lambda = 0$ and 9×9 patches.



(a) Segmented lesion

(b) 1 × 1

(c) 3 × 3

(d) 5 × 5

(e) 11 × 11

(f) 17 × 17

Figure 4.8: Illustrative example of the effect of varying the window size on a lesion with a complex colour distribution. Using a 1 × 1 window does not incorporate information from neighbouring pixels, and thus overfit the skin reflectance. As the window size was increased, less skin reflectance was captured, and more spatially-relevant colour regions were observed. In this example we used $\lambda = 0$ and seven clusters.

inspection. They were chosen such that we could compare against the "base case" (one cluster), as well as more complex reconstructions while noticing that lesion images mostly comprise a few colours but have pixels saturated with lighting artefacts. There are at most six colours observed using a dermatoscope (see Section 2.2.3), however not all of these colours are observable using standard cameras. We therefore chose five colours as our upper bound for reconstruction.

Using this colour complexity analysis framework, we constructed six HLIFs that characterize the complexity of a lesion's colour distribution. Each feature calculation can be done following the execution of the framework outlined in Section 4.3.1.

HLIFs for Quantifying Reconstruction Error

The first set of HLIFs treats the original lesion as the "ground truth". It compares the relative reconstruction errors in CIE $L^*a^*b^*$ space using one, two, and five colour reconstructions. If there is little difference between these reconstructions, it can be concluded that the colour distribution is simple; conversely, larger differences indicate more complex colour spaces.

The final formulation of these two HLIFs is as follows:

$$f_1^C = \frac{RMSD(I_{Lab}, R(I_{Lab}, 2))}{RMSD(I_{Lab}, R(I_{Lab}, 1))} \tag{4.11}$$

$$f_2^C = \frac{RMSD(I_{Lab}, R(I_{Lab}, 5))}{RMSD(I_{Lab}, R(I_{Lab}, 1))} \tag{4.12}$$

where $RMSD(I_1, I_2)$ is the root mean squared difference between images I_2 and I_2, I_{Lab} is the lesion in CIE $L^*a^*b^*$ space, and $R(I, r)$ is the colour reconstruction of image I using r colour patches. These HLIFs represent the relative amounts of reconstruction error between one-vs-two and one-vs-five colour patches, thus capturing the complexity of the colour distribution.

Figure 4.9 depicts an example of these HLIFs. Reconstruction error using one colour (i.e,. $RMSD(I_{Lab}, R(I_{Lab}, 1))$) will be large since one cluster is insufficient to reconstruct the complex colour distribution of the original lesion. The reconstruction error decreases somewhat with the two-colour reconstruction, although much of the red and pink pigmentation is still not present. The error is substantially decreased with the five-colour reconstruction, which successfully reconstructs the tan, pink, red, and dark pigmentations. The rate of reconstruction error over the number of clusters is quantified using f_1^C and f_2^C.

HLIFs for Quantifying Colour Complexity Evolution

The second set of HLIFs quantifies the amount of colour complexity by comparing the evolution of the colour distribution across reconstructions with varying numbers of colour clusters. Figure 4.7 shows the effect of increasing the number of colour clusters for a lesion with a very complex colour distribution. In this example, underlying colour patterns emerge when reconstructing with more colours, whereas a simple lesion might be accurately represented by only a single colour. To capture this information, we computed the mean "difference" between reconstruction using one, two, and five colours. This was computed using the mean ℓ^2 difference between two reconstructions in the CIE $L^*a^*b^*$ space, resulting in a value that portrays the perceptual difference between the reconstructions.

The final formulation of these two HLIFs is as follows:

$$f_3^C = \frac{1}{N}||R(I_{Lab}, 5) - R(I_{Lab}, 1)|| \tag{4.13}$$

$$f_4^C = \frac{1}{N}||R(I_{Lab}, 5) - R(I_{Lab}, 2)|| \tag{4.14}$$

where N is the number of pixels in the lesion, $||\cdot||$ is the ℓ^2 norm (i.e., Euclidean distance), and $R(I, r)$ is the reconstruction of image I with r colour clusters as in Equation 4.11 and Equation 4.12.

Figure 4.9 depicts an example of these HLIFs. Treating the one-colour reconstruction as the "base case", there is drastic colour evolution using two- and five-colour reconstructions. For example, the two dominant pigments are tan and black, both of which are reconstructed using two clusters. The less dominant red and pink pigments are reconstructed using five clusters. This evolution is successfully modeled by the HLIFs.

The third and final set of HLIFs compares the lesion's colour *signatures* across different numbers of clusters. The Earth Mover's Distance (EMD) is a very appropriate tool for quantifying this information (see Section 2.6.4). Recall from Section 2.6.4 that a signature is a cluster representation of the point distribution, where the number of points belonging to each cluster is stored. When comparing two signatures, EMD does not require these signatures to be the same size. This is an important property for our use of colour comparison.

The EMD calculates how much "work" is needed to transform one signature into another by considering the distance and mass when moving a point from one cluster to another. In our case, the distance is the Euclidean distance in CIE $L^*a^*b^*$ space, and the mass is the number of points belonging to a particular cluster after the deterministic k-means procedure.

The final formulation of these two HLIFs is as follows:

$$f_5^C = EMD(S_1, S_2) \tag{4.15}$$

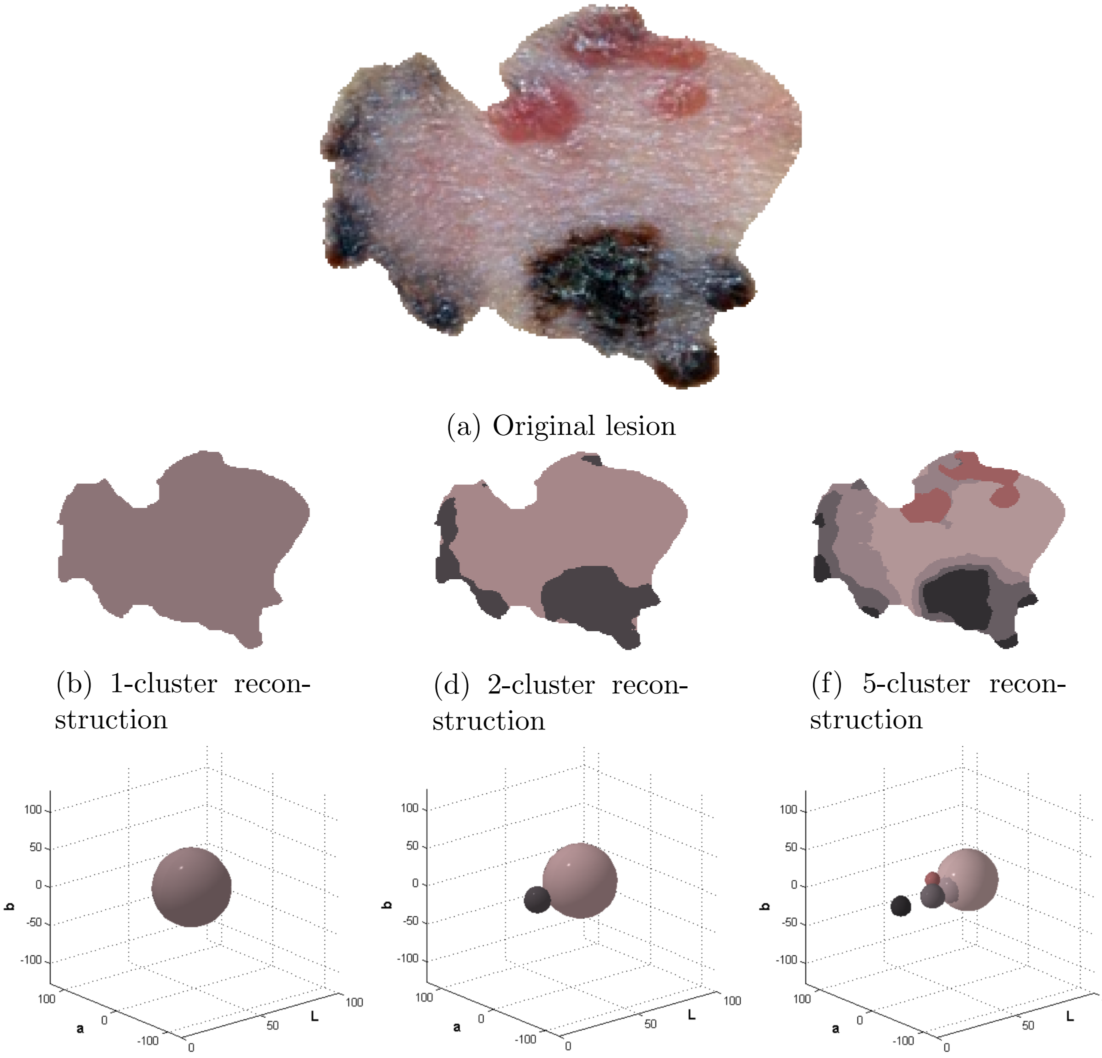
$$f_6^C = EMD(S_2, S_5) \tag{4.16}$$

where S_k is the signature using k colour clusters.

Figure 4.9 depicts an example of these HLIFs. The one-cluster signature (Figure 4.9c) is a single cluster that approximates the average lesion colour. The two-cluster signature (Figure 4.9e) picked up some of the dark pigmentation, effectively creating a smaller black cluster at a perceptual distance away from the original cluster. The five-cluster signature (Figure 4.9g) picked up the red pigmentation, as well as some of the more subtle pink pigments. It can be observed that a non-trivial amount of "work" is required to transform the one-cluster signature to the two-cluster signature by "transporting" the tan pigment to the dark pigment, and that even more work is required to transform the two-cluster signature to the five-cluster signature by creating the red and subtle pink pigments. This is indicative of a lesion with large colour complexity.

4.4 Summary

This chapter has presented 10 HLIFs for describing the asymmetry, border irregularity, and colour complexity of a skin lesion. These HLIFs have been designed with two primary criteria in mind. First, each feature has been explicitly designed to model what a dermatologist would observe when trying to identify a characteristic (e.g., colour asymmetry). Second, visual output from each feature has been given that is intuitive and describes the high-level characteristic. In the next chapter, we analyse this feature set along with a recent low-level feature set to determine the accuracy and effective nature of these features under a standard classification model.



(a) Original lesion

(b) 1-cluster reconstruction

(d) 2-cluster reconstruction

(f) 5-cluster reconstruction

(c) 1-cluster signature

(e) 2-cluster signature

(g) 5-cluster signature

Figure 4.9: Example of $\{f_1^C, \ldots, f_6^C\}$ on a superficial spreading melanoma with a complex colour distribution. Figures (b) (d) (f) are colour reconstructions using the proposed colour complexity analysis framework from Section 4.3.1, and Figures (c) (e) (g) are the associated clusters in CIE $L^*a^*b^*$ space. The size of each sphere indicates the number of pixels belonging to that cluster. The colour complexity is apparent when analysing the change in reconstructions and colour signatures as the number of clusters increases. In this case, $f_1^C = 0.682$, $f_2^C = 0.480$, $f_3^C = 0.189$, $f_4^C = 0.095$, $f_5^C = 125.3$, $f_6^C = 73.9$.

Chapter 5

Experimental Results

This chapter presents the experimental evaluation of the HLIFs proposed in Chapter 4. This feature set is analysed with and against a recent low-level feature set modeled according to the ABCD rule [27]. This low-level feature set contains 52 features that are characterised as "low level" according to the definition given in Section 3.2. The final proposed feature set is the combined set of HLIFs and low-level features. Analysis is performed in two manners. First, a simple classification scheme is presented that highlights the efficiency of the feature space rather than the classification scheme. Second, the features are statistically analysed independent of classification using our data set. Observations and limitations of the experimental results are discussed.

5.1 Data

We collected 206 images of skin lesion, which were obtained using standard consumer-grade cameras in varying and unconstrained environmental conditions. These images were extracted from the publicly available online databases Dermatology Information System [51] and DermQuest [52]. Of these images, 119 are melanomas, and 87 are not melanoma. Each image contains a single lesion of interest.

5.2 Experimental Setup

For each image, the lesion was manually segmented to provide an "ideal" segmentation for feature extraction. That is, we wished to analyse the feature extraction performance irrespective of an automatic segmentation's accuracy. We rendered the images rotation- and scale-invariant by performing the following preprocessing step: prior to feature extraction, the image was rotated so that the lesion's major axis was parallel to the horizontal axis, and the lesion fit within a 200 × 200 bounding box while maintaining the original aspect ratio. The decision support workflow was implemented in MATLAB.

5.2.1 Preprocessing

We applied an illumination correction algorithm [23]. Briefly, the illumination correction uses a Markov Chain Monte Carlo (MCMC) approach to estimate a non-parametric illumination model of the healthy skin. This model is used as a prior to fit a quadratic surface to the pixels. Finally, this quadratic surface is applied to the computed reflectance map of the image to correct for the lighting variation contributing to the non-uniform skin surface reflection.

5.2.2 Feature Extraction

Following the illumination correction, the features presented in Chapter 4 were extracted as well as the feature set proposed in [27], which is the most recent full ABCD feature set designed for standard camera images of pigmented skin lesions, to the best of the authors' knowledge. For simplicity of discussion and analysis, the following notation is used throughout this section:

- S_L – set of 52 low-level features describing ABCD proposed by [27].
- S_H – set of 10 HLIFs presented in Section 4.
- S_T – set of 62 features obtained by appending S_L and S_H (i.e., $S_T = S_L \bigcup S_H$).

Prior to passing the feature vectors into the classification stage (see below), *feature scaling* was performed according to the following normalization formula:

$$f_i^* = \frac{f_i - \mu_f}{\sigma_f} \tag{5.1}$$

where f_i^* is the normalized feature value, f_i is the feature score for the i^{th} datum, and μ_f and σ_f are the mean and standard deviation over all computed feature scores f_i. This formulation transforms the data such that each feature exhibits zero-mean and unit standard deviation (and variance) across the data set. It has been shown that scaling feature vectors eases numerical difficulties in SVM's optimization, and may result in better classification performance [53]. Furthermore, more relevant to this thesis' work, this simplifies the task of determining the feature score's significance, as each feature score distribution abides by a normal (Gaussian) distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$. For example, a feature score of $f_i = 2$ signifies that it is two standard deviations away from the mean feature score exhibited by the data set, representing that it is larger than roughly 98% of the rest of the data.

5.2.3 Classification

In machine learning, "classification" relates to a body of methods that takes as input a feature vector and outputs a predicted class (e.g., malignant or benign). In *supervised learning*, the ground-truth class of each datum is known. The classifier is usually trained (i.e., fit) using a subset of this data and tested on the other part of the data. This ensures that the classifier is able to generalise to new instances rather than strictly conforming to the training data (called *overfitting*).

Choice of Classifier

There are many different flavours of classifiers. For our purposes we used a soft margin SVM classifier due to its widely regarded robustness and simplicity (see Section 2.6.6 for details on SVM). Furthermore, we used a linear SVM to emphasize the degree of linear separability of the data in the feature space rather than the performance of a complex classifier (which is important in decision support systems, but is out of the scope of the feature extraction stage). Good accuracy can therefore be attributed to the feature extraction algorithm's ability to project the data into a separable feature space. We used the LIBSVM implementation for our experiments [54].

SVM Parameter Optimization

There are two parameters that influence the linear soft margin SVM optimization, denoted by c and w_i in LIBSVM. The parameter c (also commonly referred to as the *box constraint*) represents the cost of a misclassification during training. Thus, high c values put great emphasis on misclassified training samples. The parameter w_i assigns a weighting to c

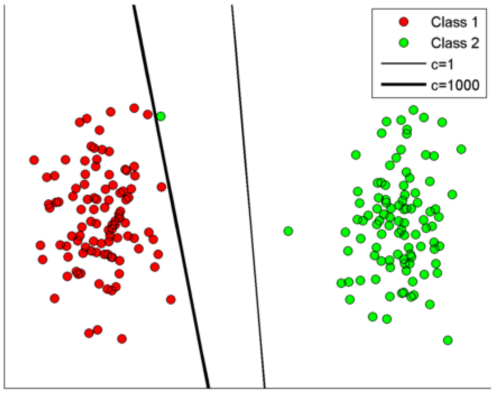


Figure 5.1: Effect of varying cost parameter c using LIBSVM. Notice how as c gets larger, more emphasis is placed on the green outlier, although it is visually apparent that some sort of error was introduced when calculating that data point. Lower values of c are robust against this phenomenon.

for class i. This weight is important when the two classes have unequal amounts of data. Mathematically, Equation 2.13 is modified to become:

$$\min_{w,s,b} \left\{ \frac{1}{2}||\mathbf{w}||^2 + \sum_{i=1}^{n} C_i s_i \right\} \text{ subject to } y_i(\mathbf{w} \cdot \mathbf{x}_i - b) \geq 1 \tag{5.2}$$

where $\mathbf{w}$ is a vector orthogonal to the separating hyperplane, $C_i = w_i c$ is a class-specific relative weighting, s_i are the slack variables associated with classification error, and $y_i \in \{-1, 1\}$ is a binary class label. In noisy data, usually a smaller value for c is desired since it allows the classifier to be robust against misclassification from the noise. An illustrative example of the effect of parameter c is provided in Figure 5.1.

To find an accurate SVM hyperplane for the data, we optimized these parameters in accordance with the LIBSVM authors' recommendations [53]. In particular, we performed a geometric grid search over the values $2^{-6} \leq c \leq 2^9$ and $2^{-2} \leq w \leq 2^2$, at each step multiplying either c or w by 2. For each pair of parameter values (c_i, w_i), we calculated the average F_β [55] across 50 independent cross-validation trials, using a random 80%/20% data split for training/testing. F_β is the weighted harmonic mean between recall and precision. Mathematically:

$$F_\beta = \frac{\text{precision} \cdot \text{recall}}{\beta^2 \cdot \text{precision} + \text{recall}} \tag{5.3}$$

where

$$\text{precision} = \frac{TP}{TP + FP} \tag{5.4}$$

$$\text{recall} = \frac{TP}{TP + FN} \tag{5.5}$$

where TP, FP, FN are the number of true positive (correct malignant detection), false positive (false malignant detection), and false negative (false benign detection) cases. In this formula, recall is weighted β-times as important as precision. See Figure 5.2 for a graphical depiction of F_β for different values of precision and recall. If the average F_β using (c_i, w_i) was greater than the previous maximum F_β, (c_i, w_i) were stored.

After evaluating the entire geometric grid, a more fine-grained search was done on the parameter values $(\hat{c}, \hat{w})$ that exhibited the largest average F_β. This stage was very similar as the previous, except the grid space was chosen according to $\frac{1}{2}\hat{c} \leq c \leq 2\hat{c}$ and $\frac{1}{2}\hat{w} \leq w \leq 2\hat{w}$, where $c_{i+1} = 2^{0.15}c_i$ and $w_{i+1} = 2^{0.5}w_i$. The parameters (c^*, w^*) that yielded the maximum average F_β were used as the final classification parameters.

Evaluating Success

Due to the small nature of our data set relative to the highly dimensional feature spaces, we used the leave-one-out cross-validation (LOO CV) strategy for evaluating the success of the classification. LOO CV is useful when dealing with this problem – that is, evaluating the classifier's ability to generalise under limited amounts of data. In particular, for each datum's feature vector $\mathbf{f_k} \in S = \{\mathbf{f_1}, \ldots, \mathbf{f_n}\}$, the SVM classifier was trained on $S \setminus \mathbf{f_k}$ and tested on $\mathbf{f_k}$, yielding a binary result: pass or fail. For a data set with n elements, this strategy resulted in n independent training and testing phases, of which the total error was determined by the total number of incorrect predictions divided by n.

Classification accuracy metrics are summarized in Table 5.1. These are discussed in detail in the following sections.

5.3 Results

Results using the 206-image data set are presented below. The thesis contribution is analysed in two ways: using the results of the HLIF set S_H on its own, and analysing the effect of concatenating S_H to the state-of-the-art low-level feature set S_L. The following analyses were performed. First, results using the described classification scheme are presented.

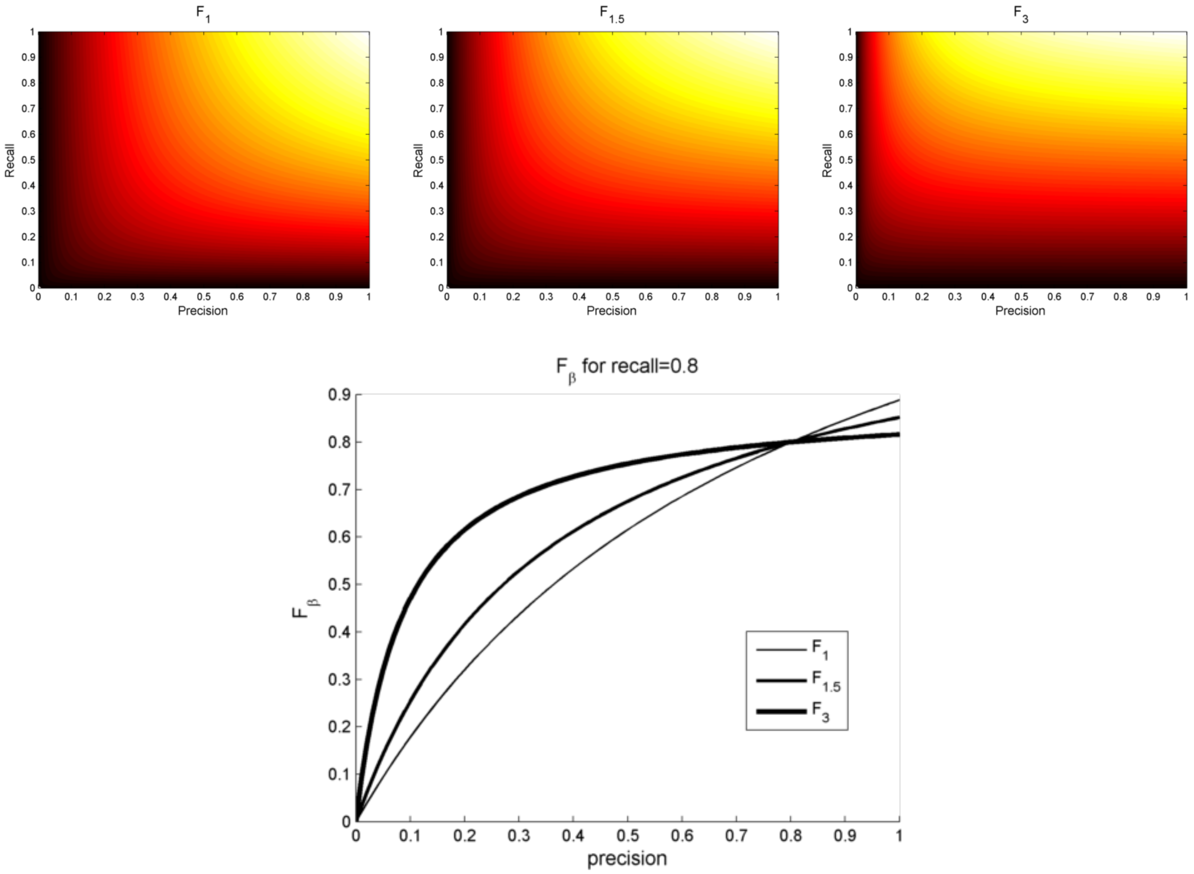


Figure 5.2: Effect of precision and recall on F_β for different values of β. Brighter colours denote large F_β scores. Notice how larger β yields a criterion that favours high recall more than high precision. This is especially apparent in the plot, where F_β elicits a stronger response to a high recall rate for most precision values as β gets larger.

Table 5.1: Definitions of accuracy metrics.

Metric	Formula	Definition
Sensitivity	$\frac{TP}{TP+FN}$	Percentage of correctly identified melanoma cases.
Specificity	$\frac{TN}{TN+FP}$	Percentage of correctly identified non-melanoma cases.
Precision	$\frac{TP}{TP+FP}$	Percentage of identified melanoma cases being correct.
Recall	$\frac{TP}{TP+FN}$	Percentage of correctly identified melanoma cases.
Accuracy	$\frac{TP+TN}{TP+FN+TN+FP}$	Percentage of correctly identified cases.
F_β	$\frac{\text{precision}\cdot\text{recall}}{\beta^2\cdot\text{precision}+\text{recall}}$	Weighted harmonic average of precision and recall.

Second, a statistical analysis of the extracted features independent of any classification scheme is presented. Limitations of the results are also discussed.

5.3.1 Classification Results

The results using the classification scheme presented in Section 5.2.3 with the feature sets S_L, S_H, and $S_T = S_L \bigcup S_H$ (see Section 5.2.2) are summarized in Table 5.2. For comparison purposes, SVM parameter selection was performed using three different values for F_β ($\beta = \{1, 1.25, 1.5\}$). Note that the SVM parameter selection strategy is not a deterministic one since a random 80%/20% train/test split is chosen for each cross-validation trial. The sensitivity and specificity scores were therefore obtained over 10 independent classification runs. That is, the SVM parameter selection and classification scheme were executed in 10 separate instances, each yielding potentially different SVM parameters due to the stochastic nature of parameter selection. The mean and standard deviation of the accuracy metrics across the 10 trials were used to show the consistency of the results.

The best mean results in Table 5.2 are bolded for each metric (i.e., sensitivity, specificity, accuracy) and for each F_β. The following observations can be made from the results.

General Accuracy Patterns

S_H consistently attains slightly higher sensitivity metrics than the other two feature sets across each F_β (92.52%, 96.22%, 96.64% for $F_1, F_{1.25}, F_{1.5}$ respectively). However, its speci-

[1] μ mean, σ standard deviation

Table 5.2: Comparing classification results of different feature sets over 10 cross-validation trials. See Section 5.2.2 for feature set descriptions.

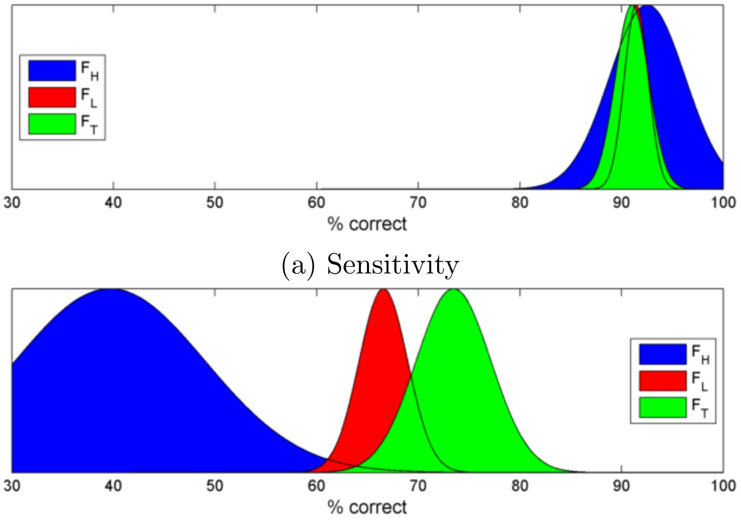
Optimization Function (F_β)	Feature set (#)	Sensitivity (%)		Specificity (%)		Accuracy (%)	
		μ^1	σ^1	μ	σ	μ	σ
F_1	S_L (52)	91.43	1.18	66.55	2.39	80.92	0.65
	S_H (10)	**92.52**	3.75	39.66	9.34	70.19	2.14
	S_T (62)	91.01	1.64	**73.45**	3.69	**83.59**	1.14
$F_{1.25}$	S_L (52)	92.94	1.54	64.48	1.38	80.92	0.64
	S_H (10)	**96.22**	1.78	32.64	2.61	69.37	0.28
	S_T (62)	92.52	1.22	**66.09**	2.78	**81.36**	1.15
$F_{1.5}$	S_L (52)	94.37	1.05	57.59	5.24	78.83	1.93
	S_H (10)	**96.64**	0.56	31.49	2.18	69.13	0.92
	S_T (62)	92.94	1.95	**65.06**	7.21	**81.17**	2.03

ficity is usually low compared to the other two feature sets. This is examined further below. Conversely, S_T consistently attains the highest specificity and accuracy metrics of all the feature sets. This indicates that the concatenation of the HLIF set S_H with the low-level feature set S_L is a valid one, yielding high classification results.

Figure 5.3 graphically portrays how the mean and standard deviation of sensitivity and specificity change across feature sets (assuming β=1 for illustrative purposes). Classification using S_T (green) exhibits approximately a 10% increase in specificity over S_L (red), while attaining high sensitivity with a small standard deviation. This alludes to S_T's capability of attaining high performance scores with high test-retest reliability. S_H attains high sensitivity, but has a relatively large standard deviation indicating moderate test-retest reliability. Adding the low-level features helps the test-retest reliability, however theoretically a larger data set should significantly help this shortcoming since the HLIF values would be more indicative of the population distribution. Since HLIFs are designed to model human-observable characteristics, lots of data is required to "learn" such high-level characteristics. This is discussed further in Section 5.3.4.

Analysing the Results of S_H

Interesting metrics are observed for the S_H feature set. It attains very high sensitivity (consistently above 90%) but conversely low specificity (consistently below 40%). This behaviour is due to the effects exhibited by using a linear classifier in a low dimensional



(a) Sensitivity

(b) Specificity

Figure 5.3: Graphical representation of classification accuracy from the results presented in Table 5.2 (using F_1). Notice how appending S_H to S_L increases the mean sensitivity while maintaining similar sensitivity standard deviation in S_T, therefore increasing performance while maintaining test-retest reliability.

feature space on a small data set. This can be explained by considering the following case. Assume that the classifier predicts everything as malignant (i.e., $\forall \mathbf{f_i},\ \hat{y}(\mathbf{f_i}) = 1$). Accordingly, precision and recall values become as follows:

$$
\begin{aligned}
& TN = FN = 0 \\
\Rightarrow & prec = \frac{TP}{TP + FP} = \frac{n_m}{n} \\
\Rightarrow & rec \ \ = \frac{TP}{TP + 0} = 1
\end{aligned}
$$

where n_m is the number of malignant cases in the data and n is the total number of data. This classification translates to a perfect recall score, and a precision score proportional to the percentage of malignant cases in the data. Now, consider the following facts. In our case, 57% of the data is malignant melanoma. Melanoma can take on many different shapes and colours (see Chapter 2). Remember that these images are uncalibrated images obtained in unconstrained environments. It is difficult for a small number of features to successfully separate the classes in such a setting. Our data set is relatively small. All of these factors contribute to the fact that it would be difficult for the linear classifier to

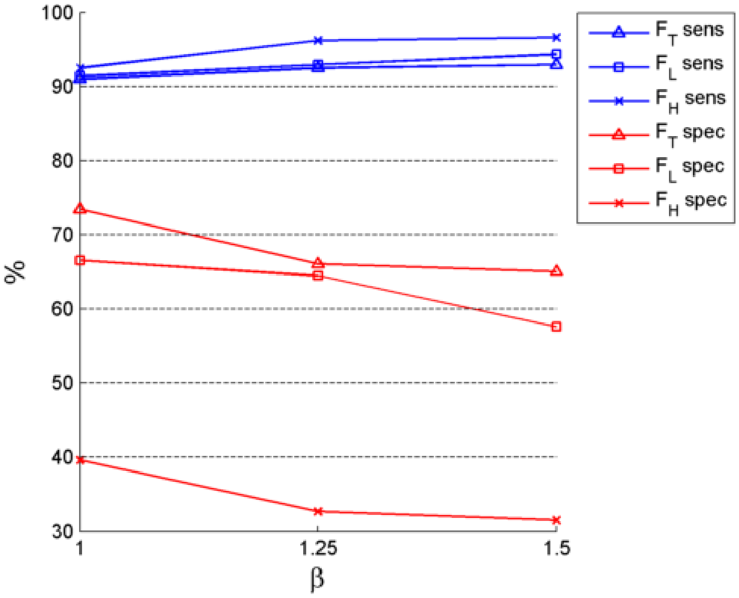


Figure 5.4: Effect of β in F_β on mean sensitivity and specificity from the results presented in Table 5.2. Blue lines are sensitivity curves, and red lines are specificity curves. Notice how sensitivity rises with increased β, whereas specificity drops.

attain a higher F_β without sacrificing recall in a low-dimensional space. This problem can theoretically be abated with a larger data set that is more representative of the population distributions, as discussed further in Section 5.3.4.

Effect of F_β on Classification Results

The choice of F_β as an objective function for SVM parameter selection affects the final classification results. Remember that a higher β weighs recall higher than precision during optimization, effectively varying the relative "importance" of recall. In fact, a value of β translates to recall being β-times as "important" as precision by the definition of the weighted harmonic mean. Since the mathematical formulation of recall is the same as that of sensitivity, it makes sense that as β increases, sensitivity increases and specificity decreases. This is indeed observed in Table 5.2 for all three feature sets. This pattern is shown graphically in Figure 5.4. This parameter β can be tuned according to the user's preference regarding false positive and false negative rates.

Table 5.3: Welch's two-sample t-test on the set of HLIFs

HLIF	f_1^A	f_2^A	f_1^B	f_2^B	f_1^C	f_2^C	f_3^C	f_4^C	f_5^C	f_6^C
p**-value**	<0.001	0.036	<0.001	<0.001	0.004	0.039	0.13	0.026	0.048	0.002

5.3.2 Statistical Evaluation of the Feature Space

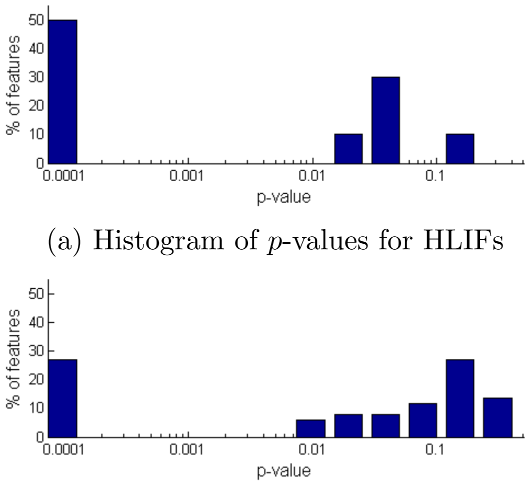
The feature space can be statistically analysed for class separability. This provides classifier-independent measures of the performance of each feature at identifying melanoma. Although these tests hold their own set of assumptions, conclusions can be drawn on each feature independently, thus contributing to a more holistic analysis of the proposed features. The metrics used to evaluate the feature space are described here, along with the computed results.

Two-sample t-test

A two-sample t-test tests the statistical difference of two distributions. It assumes that the classes are normally distributed with unknown but unequal variances according to $\mathcal{N}(\mu, \sigma_i^2)$, where μ is the population mean and σ_i^2 is the class variance. The feature score normalization step (see Section 5.2.2) transforms the feature scores to a zero-mean unit-variance distribution. Specifically, given sample data from two classes, Welch's two-sample t-test seeks to reject the null hypothesis H_0 that two class sample distributions come from the same population distribution.

Table 5.3 shows the results of running Welch's two-sample t-test for each HLIF. A low p-value indicates that a particular feature differentiates between the two classes (malignant and benign) very well, under the assumption of normality. Three of the features exhibit p-values less than 0.001, indicating that they are good feature descriptions. Further, only one feature (f_3^C) exhibits a p-value greater than 0.10. However, this feature is kept, since it adds information to f_4^C, and 0.10 is not a terrible score when dealing with the noisy unconstrained environment of skin lesions in standard camera images.

The normalized histograms of p-values for S_L and S_H using Welch's two-sample t-test are shown in Figure 5.5. A larger percentage of features from S_H exhibit low p-values compared to those from S_L. for example, 50% of the features from S_H exhibited p-values less than 0.001, whereas only 28% of the features from S_L exhibited such low p-values. These results alludes to the discriminative power of HLIFs by nature for problems that involve improving intuitive identification.



(a) Histogram of p-values for HLIFs

(b) Histogram of p-values for low-level features

Figure 5.5: Comparison of p-values using Welch's two-sample t-test on each of HLIFs and low-level features from [27]. The x-axis is log-scale, indicative of p-value importance. Notice how a much higher percentage of HLIFs exhibit small p-values than the low-level features, alluding to the discriminating power of HLIFs.

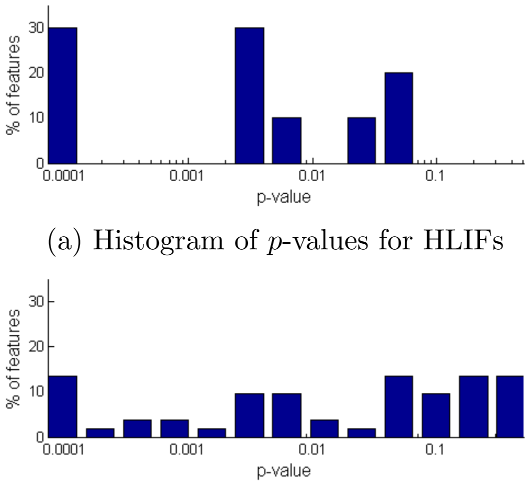
Table 5.4: Mann-Whitney U test on the set of HLIFs

HLIF	f_1^A	f_2^A	f_1^B	f_2^B	f_1^C	f_2^C	f_3^C	f_4^C	f_5^C	f_6^C
p-value	<0.001	0.003	<0.001	<0.001	0.003	0.03	0.061	0.007	0.072	0.004

Mann-Whitney U Test

Another test of data separability is the non-parametric Mann-Whitney U test [56]. Given sample data from two classes, the Mann-Whitney U test seeks to reject the null hypothesis H_0 that two class sample distributions come from the same population distribution. It assumes that the classes are *not* normally distributed, and that the samples were obtained independently.

Table 5.4 shows a similar story as the t-test results: HLIFs are good at discriminating between the classes for which they are designed. To further emphasize this point, Figure 5.6 shows that a much larger percentage of HLIFs attain lower p-values than low-level features. These results are consistent with the results obtained in Section 5.3.2, providing further evidence that HLIFs are very discriminatory especially compared to low-level features.



(a) Histogram of p-values for HLIFs

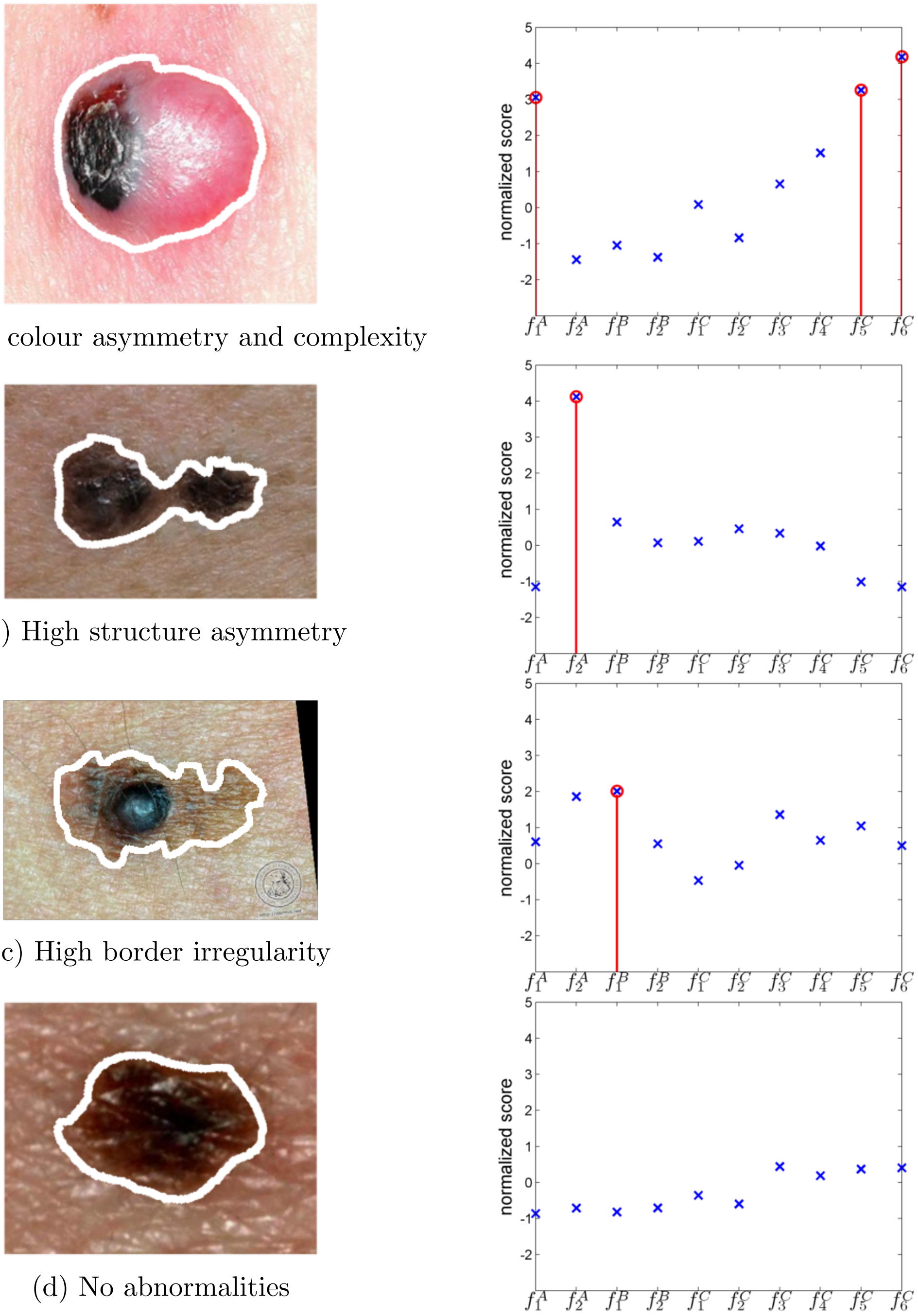
(b) Histogram of p-values for low-level features

Figure 5.6: Comparison of p-values using the non-parametric Mann-Whitney U test on each of HLIFs and low-level features from [27]. The x-axis is log-scale, indicative of p-value importance. Much like the results of the t-test, a larger percentage of HLIFs exhibit small p-values than the low-level features, alluding to the discriminating power of HLIFs.

5.3.3 Providing Intuitive Rationale

It stands to reason that a doctor is more likely to trust a computer-generated malignancy prediction if intuitive rationale is provided along with the predicted label. Although it is out of the scope of this thesis, this rationale can also be used in a reinforcement learning scheme, where the user can give intuitive feedback to evolve the classifier. Furthermore, the output can be used as a learning aid to new or training dermatologists, to test their knowledge of the widely-used ABCD rule.

Figure 5.7 shows the 10 HLIF scores for a set of images, along with the qualitative interpretation of the feature calculations (e.g., "High colour asymmetry"). The features had been normalized so that the significance of a feature calculation can be easily interpreted by a feature score's number of standard deviations from the sample mean. Figure 5.8 provides an example interface for intuitive visualization of the colour asymmetry.


normalized score
5
4
3
2
1
0
-1
-2
f_1^A f_2^A f_1^B f_2^B f_1^C f_2^C f_3^C f_4^C f_5^C f_6^C

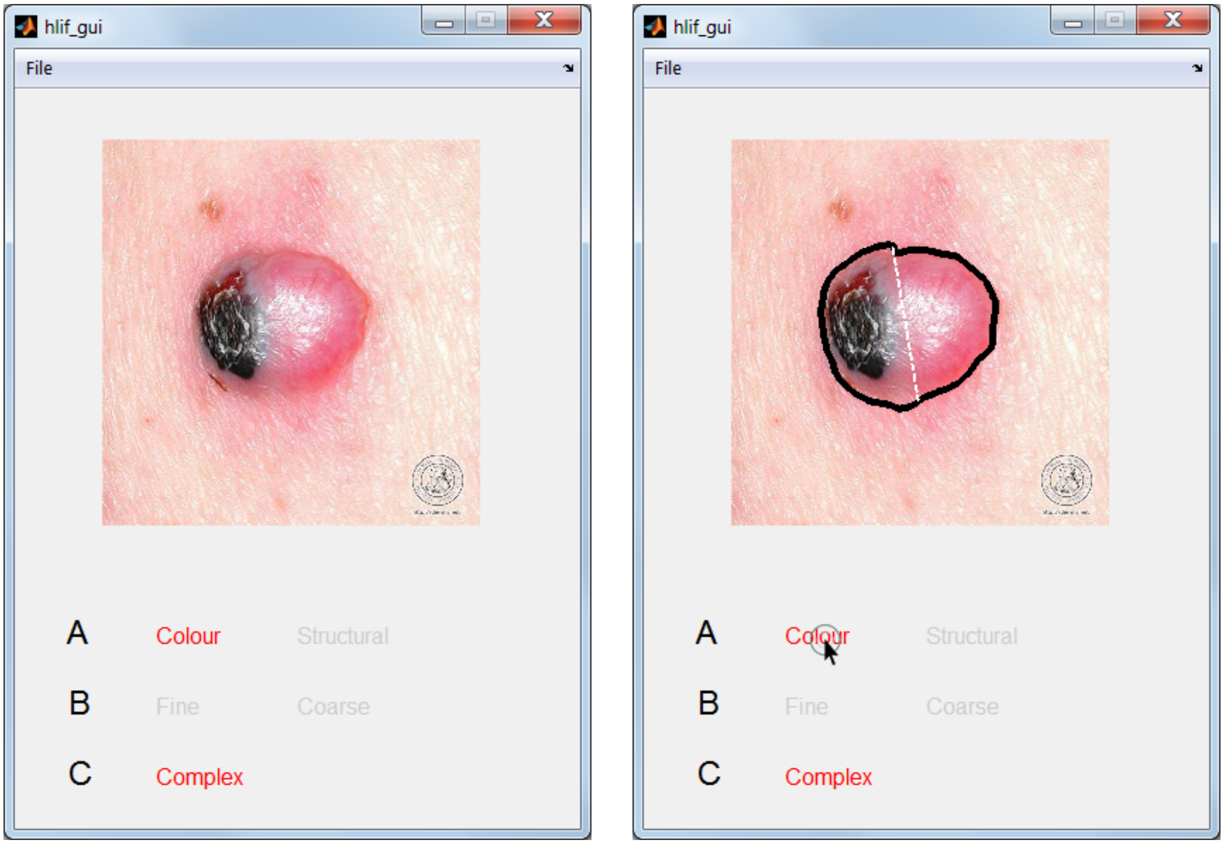
(a) High colour asymmetry and complexity

(b) High structure asymmetry

(c) High border irregularity

(d) No abnormalities

Figure 5.7: Example HLIFs scores. Abnormally high feature scores are circled in red.



(a) Lesion analysis results

(b) Intuitive visualization of colour asymmetry

Figure 5.8: Intuitive visualization for the case presented in Figure 5.7a. Upon analysing the image, the interface indicates that there is apparent colour asymmetry and complex colour patterns by highlighting the relevant ABCD terms. When the user clicks on “Colour”, an overlay is shown to provide intuitive justification for the claim.

5.3.4 Discussion and Limitations

One over-arching conclusion can be drawn from the experimental results: HLIFs capture relevant information for melanoma detection. We showed that a small set of HLIFs can increase classification performance when combined with a large set of low-level features, and that single HLIFs, if designed properly, are more effective at data separation than single low-level features. The performance of the HLIFs themselves are not as powerful as we would have hoped, however they do present one significant advantage over large low-level feature sets: a small HLIF set requires much less data to adequately populate the multi-dimensional feature space. A larger data set may allow the HLIF set to perform better, as the classifier could be trained on more representative class distributions. In its current state, insufficient data leads to suboptimal results.

A large hindrance of the current state of skin cancer detection research is the limited amount of data available to the scientific community. Dermatologists may take pictures of skin lesions, but restrict them to within their clinic, due perhaps to either privacy or commercialization concerns. In order to ensure robust models and statistical validity, much larger data sets must be accumulated for training and testing these decision support systems. This is especially the case since the images are obtained in unconstrained environments, leading to extremely large variations in acquired data.

Another unfortunate by-product of the current data collection methods is that most images presented with a final diagnosis are of late-stage melanoma. Melanoma patients' prognosis is highly correlated with the stage in which it is identified (and excised). Although there is merit in mid- to late-stage melanoma diagnosis, a large emphasis should be placed on early-stage diagnosis for ensuring better survival rates. Again, data collection needs to precede the validation of models for the systems to be accurate.

5.4 Summary

This chapter has presented empirical evidence supporting the effective nature of HLIFs. The features presented in Chapter 4 were analysed in a classification framework as well as a statistical classification-independent framework. These analyses have shown that, in combination with low-level features, the small set of HLIFs generates promising results. Furthermore, it was shown that individual HLIFs tend to represent meaningful information on a greater level than individual low-level features, leading to better data separation with small feature sets.

Chapter 6

Conclusions

6.1 Summary

This thesis has presented a framework for designing high-level intuitive features (HLIFs), and has proposed a set of HLIFs for quantifying skin lesion characteristics with the goal of detecting melanoma. HLIFs are feature calculations that have been meticulously modeled to describe some human-observable characteristic, and of which rationale can be relayed to the user in some intuitive (perhaps visual) manner. It was shown in Section 5.3 that skin lesion classification accuracy was improved when concatenating a small set of HLIFs to a state-of-the-art low-level feature set. Individual HLIFs were shown to have more statistical significance with respect to separating the data than individual low-level features in Section 5.3.2. Furthermore, the addition of HLIFs enabled the provision of intuitive rationale for the predicted malignancy, as shown in Section 5.3.3.

Chapter 3 presented a framework for designing HLIFs for a classification problem in which the automatic classification scheme mimics or enhances human interpretation. Skin cancer detection is such a case. That is, since dermatologists visually analyse skin lesions using the ABCD rule, HLIFs can be used to automatically classify images of skin lesions.

Chapter 4 presented 10 HLIFs for describing asymmetry, border irregularity, and colour variations of skin lesion images obtained with standard consumer-grade cameras using the HLIF framework. The design of these features were inspired by the ABCD rule by which many dermatologists visually analyse skin lesions for malignancy. In total, we proposed two HLIFs for describing colour and structure asymmetry, two HLIFs for describing fine and coarse border irregularity, and six HLIFs for describing colour complexity.

Chapter 5 provided an experimental setup and the subsequent evaluation of the HLIFs from Chapter 4 as well as state-of-the-art low-level features using a data set amalgamated from public image repositories. We evaluated the efficiency of the HLIF set separately as well as its effect when concatenated with the low-level feature set. The feature sets were analysed using a linear SVM classifier, as well as using statistical tests of the feature space independently of a classifier. Our experimental results indicated that HLIFs increased classification accuracy of low-level features, HLIFs were individually better at separating image data properly than the low-level features, and HLIFs provided means for conveying diagnostic rationale to the user.

6.2 Future Work

It is the hope that this work may lead to more clinically-viable decision support systems with the aid of powerful HLIF sets. As such, we have identified several areas of future work that would benefit the skin cancer detection research.

6.2.1 Melanin and Hemoglobin Characterization

There has been a recent rise in melanin and hemoglobin information extraction techniques using standard camera images [57, 17, 58]. Melanin and hemoglobin concentrations can help elucidate information not immediately present in the original image. Using an accurate hemoglobin/melanin extraction model, HLIFs can be designed to capture patterns about the hemoglobin and melanin densities. Analysing the melanin information holds promise, as melanoma is characterized by a metastatic growth of melanin-producing melanocytes.

6.2.2 HLIFs for Other Melanoma Patterns

Although the ABCD rule is powerful and widely used, there are other rules which lend themselves to detecting melanoma. Among them are the Menzies method [59], seven-point checklist [60], and CASH [61]. Furthermore, there have recently been arguments that the ABCD rule is not very well-suited to detecting nodular melanoma [62]. This type of melanoma is usually characterized by a circular border (i.e., no border irregularity), and limited radial growth. A new rule, termed EFG, has been proposed to specifically detect nodular melanoma [62]. There is therefore a demand to design HLIFs according to these rules, so that more intuitive information can be captured, processed, and relayed back to the user.

6.2.3 Analysing Lesion Evolution

The spatiotemporal evolution of a lesion can play an important role in early melanoma detection. As discussed in Section 2.1.3, melanoma lesions undergo radial and vertical growth phases. The rate at which lesions grow through these phases can be an indicator of malignancy.

6.2.4 Dimensionality Reduction

Standard camera images obtained in unconstrained environments provide many technical challenges. Features must be designed carefully to be maximally invariant to changes in lighting, camera-specific RGB sensors, etc. Due to the noisy and unconstrained environment, many features are required to properly capture the intended diagnostic information. This may lead to highly dimensional spaces, as is seen in low-level feature sets [27]. Due to the lack of large data sets, this may pose a problem for classifier learning in such a sparse feature space. There is therefore a need to design and apply dimensionality reduction techniques to reduce the feature space dimensionality while not overfitting to the training data. This may yield models with better generalisability and less prone to overfitting.

6.2.5 Data

A much larger endeavour that may have the largest impact on skin lesion detection research is the systematic construction of a comprehensive data set. The current state of the research is plagued with the lack of good data. With this data shortage, it is difficult to train and analyse the decision support systems in a manner that can successfully generalise to new lesions.

References

[1] A. F. Jerant, J. T. Johnson, C. D. Sheridan, and T. J. Caffrey, “Early detection and treatment of skin cancer,” *American Family Physician*, vol. 62, no. 2, pp. 357–368, 2000.

[2] R. Lucas, T. McMichael, W. Smith, and B. Armstrong, *Solar Ultraviolet Radiation: Global burden of disease from solar ultraviolet radiation*, ser. Environmental Burden of Disease Series, A. Pruss-Ustun, H. Zeeb, C. Mathers, and M. Repacholi, Eds. World Health Organization, 2006, no. 13.

[3] H. C. Engasser and E. M. Warshaw, “Dermatoscopy use by US dermatologists: a cross-sectional survey,” *Journal of the American Academy of Dermatology*, vol. 63, no. 3, pp. 412–419, 2010.

[4] American Cancer Society, “Cancer Facts & Figures 2011,” Atlanta, GA, Tech. Rep. ACSPC-029771, 2011.

[5] B. M. Muir, “Trust between humans and machines, and the design of decision aids,” *International Journal of Man-Machine Studies*, vol. 27, pp. 527–539, 1987.

[6] B. J. Fogg, “Persuasive technology: using computers to change what we think and do,” *Ubiquity*, vol. 2002, no. December, 2002.

[7] B. Young, J. S. Lowe, A. Stevens, and J. W. Heath, *Wheater’s functional histology: a text and colour atlas*, 5th ed. Churchill Livingstone, 2006.

[8] R. R. Anderson and J. A. Parrish, “The optics of human skin,” *Journal of Investigative Dermatology*, vol. 77, pp. 13–19, 1981.

[9] Canadian Cancer Society, http://www.cancer.ca, 2013, Accessed: 12 Aug 2013.

[10] W. Stolz, A. Riemann, A. Cognetta, L. Pillet, W. Abmayr, D. Holzel, P. Bilek, F. Nachbar, M. Landthaler, and O. Braun-Falco, "ABCD rule of dermatoscopy: a new practical method for early recognition of malignant melanoma," *European Journal of Dermatology*, vol. 4, no. 7, pp. 521–527, 1994.

[11] F. Nachbar, W. Stolz, T. Merkle, A. B. 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*, vol. 30, no. 4, pp. 551–559, 1994.

[12] E. L. Psaty and A. C. Halpern, "Current and emerging technologies in melanoma diagnosis: the state of the art," *Clinics in Dermatology*, vol. 27, no. 1, pp. 35–45, 2009.

[13] M. Vestergaard, P. Macaskill, P. Holt, and S. Menzies, "Dermoscopy compared with naked eye examination for the diagnosis of primary melanoma: a meta-analysis of studies performed in a clinical setting," *British Journal of Dermatology*, vol. 159, no. 3, pp. 669–676, 2008.

[14] D. S. Rigel, J. Russak, and R. Friedman, "The evolution of melanoma diagnosis: 25 years beyond the abcds," *CA: A Cancer Journal for Clinicians*, vol. 60, no. 5, pp. 301–316, 2010.

[15] S. D. Cotton, "A non-invasive imaging system for assisting in the diagnosis of malignant melanoma," Ph.D. dissertation, University of Birmingham, UK, 1998.

[16] M. Moncrieff, S. Cotton, E. Claridge, and P. Hall, "Spectrophotometric intracutaneous analysis: a new technique for imaging pigmented skin lesions," *British Journal of Dermatology*, vol. 146, no. 3, pp. 448–457, 2002.

[17] B. D'Alessandro and A. P. Dhawan, "3-d volume reconstruction of skin lesions for melanin and blood volume estimation and lesion severity analysis," *Medical Imaging, IEEE Transactions on*, vol. 31, no. 11, pp. 2083–2092, nov 2012.

[18] Verisante Technology Inc., http://www.verisante.com/products/aura, 2013, Accessed: 08 Jul 2013.

[19] MELA Sciences, http://melasciences.com/melafind, 2013, Accessed: 08 Jul 2013.

[20] D. Huang, E. A. Swanson, C. P. Lin, J. S. Schuman, W. G. Stinson, W. Chang, M. R. Hee, T. Flotte, K. Gregory, C. A. Puliafito *et al.*, "Optical coherence tomography," *Science*, vol. 254, no. 5035, pp. 1178–1181, 1991.

[21] T. Gambichler, P. Regeniter, F. G. Bechara, A. Orlikov, R. Vasa, G. Moussa, M. Stcker, P. Altmeyer, and K. Hoffmann, "Characterization of benign and malignant melanocytic skin lesions using optical coherence tomography in vivo."

[22] T. Gambichler, V. Jaedicke, and S. Terras, "Optical coherence tomography in dermatology: technical and clinical aspects," *Archives of Dermatological Research*, vol. 303, no. 7, pp. 457–473, 2011.

[23] J. Glaister, R. Amelard, A. Wong, and D. A. Clausi, "MSIM: Multistage illumination modeling of dermatological photographs for illumination-corrected skin lesion analysis," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 7, pp. 1873–1883, 2013.

[24] R. Gonzalez and R. Woods, *Digital Image Processing*, 3rd ed. Pearson Education, 2011.

[25] I. Maglogiannis and C. N. Doukas, "Overview of advanced computer vision systems for skin lesions characterization," *Information Technology in Biomedicine, IEEE Transactions on*, vol. 13, no. 5, pp. 721–733, 2009.

[26] K. Korotkov and R. Garcia, "Computerized analysis of pigmented skin lesions: A review," *Artificial Intelligence in Medicine*, vol. 56, no. 2, pp. 69–90, 2012.

[27] P. G. Cavalcanti and J. Scharcanski, "Automated prescreening of pigmented skin lesions using standard cameras," *Computerized Medical Imaging and Graphics*, vol. 35, no. 6, pp. 481–491, 2011.

[28] J. F. Alcon, C. Ciuhu, W. Ten Kate, A. Heinrich, N. Uzunbajakava, G. Krekels, D. Siem, and G. de Haan, "Automatic imaging system with decision support for inspection of pigmented skin lesions and melanoma diagnosis," *IEEE Journal of Selected Topics in Signal Processing*, vol. 3, no. 1, pp. 14–25, 2009.

[29] L. Ballerini, R. B. Fisher, B. Aldridge, and J. Rees, "A color and texture based hierarchical K-NN approach to the classification of non-melanoma skin lesions," in *Color Medical Image Analysis*, ser. Lecture Notes in Computational Vision and Biomechanics, M. E. Celebi and G. Schaefer, Eds. Springer Netherlands, 2013, vol. 6, pp. 63–86.

[30] R. Amelard, A. Wong, and D. A. Clausi, "Extracting high-level intuitive features (HLIF) for classifying skin lesions using standard camera images," in *CRV'12: Ninth Conference on Computer and Robot Vision*, Toronto, may 2012, pp. 396–403.

[31] ——, "Extracting morphological high-level intuitive features (HLIF) for enhancing skin lesion classification," in *EMBC'12: 34th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, San Diego, 28 2012–sept. 1 2012, pp. 4458–4461.

[32] R. Amelard, J. Glaister, A. Wong, and D. A. Clausi, "Melanoma decision support using lighting-corrected intuitive feature models," in *Computer Vision Techniques for the Diagnosis of Skin Cancer*, ser. Series in BioEngineering, J. Scharcanski and M. E. Celebi, Eds. Springer, 2013.

[33] C. M. Bishop, *Pattern recognition and machine learning.* Springer New York, 2006.

[34] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the fifth annual workshop on Computational learning theory*, ser. COLT '92. New York, NY, USA: ACM, 1992, pp. 144–152.

[35] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, pp. 273–297, 1995.

[36] J. M. Tripp, A. W. Kopf, A. A. Marghoob, and R. S. Bart, "Management of dysplastic nevi: a survey of fellows of the american academy of dermatology," *Journal of the American Academy of Dermatology*, vol. 46, no. 5, pp. 674–682, 2002.

[37] T. K. Lee, M. S. Atkins, R. P. Gallagher, C. E. MacAulay, A. Coldman, and D. I. McLean, "Describing the structural shape of melanocytic lesions," in *SPIE Medical Imaging*, 1999, pp. 1170–1179.

[38] B. S. Aribisala and E. Claridge, "A border irregularity measure using a modified conditional entropy method as a malignant melanoma predictor," in *Image Analysis and Recognition*, ser. Lecture Notes in Computer Science, M. Kamel and A. Campilho, Eds. Springer Heidelberg, 2005, vol. 3656, pp. 914–921.

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

[40] CIE, "Commission internationale de l'eclairage proceedings 1931," 1932.

[41] W. D. Wright, "A re-determination of the trichromatic coefficients of the spectral colours," *Transactions of the Optical Society*, vol. 30, no. 4, pp. 141–164, 1929.

[42] J. Guild, “The colorimetric properties of the spectrum,” *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, vol. 230, pp. 149–187, 1932.

[43] J. Serra, *Image analysis and mathematical morphology.* London: Academic Press, 1982.

[44] Y. Rubner, C. Tomasi, and L. J. Guibas, “The earth mover’s distance as a metric for image retrieval,” *International Journal of Computer Vision*, vol. 40, no. 2, pp. 99–121, 2000.

[45] O. Pele and M. Werman, “Fast and robust earth mover’s distances,” in *2009 IEEE 12th International Conference on Computer vision*, 2009, pp. 460–467.

[46] T. Su and J. Dy, “A deterministic method for initializing k-means clustering,” in *Tools with Artificial Intelligence, 2004. ICTAI 2004. 16th IEEE International Conference on.* IEEE, 2004, pp. 784–786.

[47] M. E. Celebi and H. A. Kingravi, “Deterministic initialization of the k-means algorithm using hierarchical clustering,” *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 26, no. 7, p. 1250018, 2012.

[48] J. A. Wolf, J. Moreau, O. Akilov, T. Patton, J. C. English III, J. Ho, and L. K. Ferris, “Diagnostic inaccuracy of smartphone applications for melanoma detection,” *JAMA Dermatology*, vol. 149, no. 4, pp. 422–426, 2013.

[49] R. E. Bellman, *Dynamic Programming.* Dover Publications, 2003.

[50] M. E. Celebi and G. Schaefer, Eds., *Color Medical Image Analysis.*

[51] Dermatology Information System, http://www.dermis.net, 2012, Accessed: 08 Nov 2012.

[52] DermQuest, http://www.dermquest.com, 2012, Accessed: 08 Nov 2012.

[53] C.-W. Hsu, C.-C. Chang, and C.-J. Lin, “A practical guide to support vector classification,” https://www.cs.sfu.ca/people/Faculty/teaching/726/spring11/svmguide.pdf, 2010, Accessed 22 Nov 2012.

[54] C.-C. Chang and C.-J. Lin, “LIBSVM: A library for support vector machines,” *ACM Transactions Intelligent System and Technology*, vol. 2, no. 3, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[55] C. van Rijsbergen, *Information Retrieval*, 2nd ed. Newton, Massachusetts: Butterworth-Heinemann, 1979.

[56] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *Annals of mathematical statistics*, vol. 18, no. 1, pp. 50–60, 1947.

[57] N. Tsumura, N. Ojima, K. Sato, M. Shiraishi, H. Shimizu, H. Nabeshima, S. Akazaki, K. Hori, and Y. Miyake, "Image-based skin color and texture analysis/synthesis by extracting hemoglobin and melanin information in the skin," *ACM Transactions on Graphics*, vol. 22, no. 3, pp. 770–779, 2003.

[58] A. Madooei, M. S. Drew, M. Sadeghi, and M. S. Atkins, "Intrinsic melanin and hemoglobin colour components for skin lesion malignancy detection," in *Medical Image Computing and Computer-Assisted Intervention MICCAI 2012*, ser. Lecture Notes in Computer Science, N. Ayache, H. Delingette, P. Golland, and K. Mori, Eds. Springer Heidelberg, 2012, vol. 7510, pp. 315–322.

[59] S. W. Menzies, C. Ingvar, K. A. Crotty, and W. H. McCarthy, "Frequency and morphologic characteristics of invasive melanomas lacking specific surface microscopic features," *Archives of Dermatology*, vol. 132, no. 10, pp. 1178–1182, 1996.

[60] 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," *Archives of Dermatology*, vol. 134, no. 12, pp. 1563–1570, 1998.

[61] J. S. Henning, S. W. Dusza, S. Q. Wang, A. A. Marghoob, H. S. Rabinovitz, D. Polsky, and A. W. Kopf, "The CASH (color, architecture, symmetry, and homogeneity) algorithm for dermoscopy," *Journal of the American Academy of Dermatology*, vol. 56, no. 1, pp. 45–52, 2007.

[62] A. J. Chamberlain, L. Fritschi, and J. W. Kelly, "Nodular melanoma: Patients' perceptions of presenting features and implications for earlier detection," *Journal of the American Academy of Dermatology*, vol. 48, no. 5, pp. 694–701, 2003.