Depth Data Improves Non-Melanoma Skin Lesion Segmentation and Diagnosis

Xiang Li

Doctor of Philosophy
School of Informatics
University of Edinburgh
2011

Abstract

Examining surface shape appearance by touching and observing a lesion from different points of view is a part of the clinical process for skin lesion diagnosis. Motivated by this, we hypothesise that surface shape embodies important information that serves to represent lesion identity and status. A new sensor, Dense Stereo Imaging System (DSIS) allows us to capture 1:1 aligned 3D surface data and 2D colour images simultaneously. This thesis investigates whether the extra surface shape appearance information, represented by features derived from the captured 3D data benefits skin lesion analysis, particularly on the tasks of segmentation and classification. In order to validate the contribution of 3D data to lesion identification, we compare the segmentations resulting from various combinations of images cues (*e.g.,* colour, depth and texture) embedded in a region-based level set segmentation method. The experiments indicate that depth is complementary to colour. Adding the 3D information reduces the error rate from 7.8% to 6.6%. For the purpose of evaluating the segmentation results, we propose a novel ground truth estimation approach that incorporates a prior pattern analysis of a set of manual segmentations. The experiments on both synthetic and real data show that this method performs favourably compared to the state of the art approach STAPLE [1] on ground truth estimation. Finally, we explore the usefulness of 3D information to non-melanoma lesion diagnosis by tests on both human and computer based classifications of five lesion types. The results provide evidence for the benefit of the additional 3D information, *i.e.,* adding the 3D-based features gives a significantly improved classification rate of 80.7% compared to only using colour features (75.3%). The three main contributions of the thesis are improved methods for lesion segmentation, non-melanoma lesion classification and lesion boundary ground-truth estimation.

Acknowledgements

I owe my deepest gratitude to my supervisor Bob Fisher, whose long-standing guidance, meticulous support and inspiring encouragement from the preliminary to the concluding level have enabled me to develop an understanding of the subject and accomplish this thesis. I have greatly respected his erudite knowledge, hard working spirit and rigorous attitude towards science as a researcher and appreciated his kind advice and solicitude as a senior. I would also like to thank my second supervisor Jonathan Rees for his support from the medical respect and his insightful comments and suggestions on my research. I can never forget his quick email inquiring the safety of my family right after the tragic earthquake took place in Si Chuan during 2008.

I have greatly appreciated my colleagues in the 'Vision Lab' and Dermatology Department and would especially like to thank Steven McDonagh and Bas Boom for their valuable advices on my thesis and Ben Aldridge and Yvonne Bisset for their time and efforts on helping me with my experiments. Also, I would like to thank my examiners Prof. Chris Williams and Prof. Ela Claridge for reviewing this work and providing valuable suggestions. I offer my regards and blessings to all of those who supported me in any respects during the completion of the thesis.

I gratefully acknowledge the funding for this study provided by the Foundation for Skin Research and School of Informatics.

Last, but not the least, I would like to thank my parents for all their moral support and encouragement.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Xiang Li*)

To my family.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Skin cancer is the most frequent type of cancer in the fair-skinned population. One in every three cancers diagnosed is a skin cancer. Different skin cancers display different symptoms and appearances. There are three main types of skin cancer, Basal Cell Carcinoma (BCC), Squamous Cell Carcinoma (SCC) and melanoma. Melanoma is a tumour derived from the melanocyte lineage. It has the highest case fatality with an instance rate in European populations of around 14 per 100,000, with a mortality rate of 2-3 per 100,000. Unless melanoma is diagnosed and excised early, it is generally untreatable. The other two tumour types, BCC and SCC are keratinocyte derived tumours. Basal cell carcinoma is the commonest human malignancy that virtually never metastasises, although is locally invasive. It is 10-20 times more common that melanoma. Squamous cell carcinoma is also keratinocyte derived, which may also metastasise but has a much lower case fatality than melanoma. All these forms of skin cancer are related to sun exposure, although the exact relationship differs for each one. Each year, about $132,000$ melanoma and between 2 and 3 million non-melanoma skin cancers occur globally [2]. In the UK, about 100,000 new cases of non-melanoma skin cancer occur every year with BCC accounting for 75% of all cases and SCC accounting for 20% [3]. In recent decades, skin cancer incidence has increased faster than that of almost all other cancers. The instance of all these tumours is increasing in pale skin populations with a doubling time of 10 years. With the dramatically increasing incidence of skin lesions, especially for the most dangerous type of skin tumour - melanoma, more attention has been focused on developing computer-aided skin lesion diagnosis systems (CSLD). A number of researchers have been working in this field since 1987 [4], when the idea of using computers in skin lesion analysis was first proposed. Many systems have been developed since then, with the ultimate goal of providing objective, consistent, quantitative and cheap diagnosis to compete with the subjective, descriptive and expensive diagnosis given by clinical experts.

In general, CSLD systems translate the knowledge of dermatologists into a computer program that applies medical image analysis techniques to the quantitative measurements of pathological alterations of human skin [5]. The key to CSLD systems is therefore these measurements, which are also called features. They are extracted to represent the morphological characteristics of a lesion surface and to classify the lesion type by investigating the correlation between them and the kind of lesion. However, because of the natural complexity and variation displayed by skin lesions, it is very difficult to extract features that are effective enough to describe the skin morphology and even to segment lesions from the surrounding skin [6]. Therefore, even after more than 20 years of effort, CSLD systems are still in the experimental stages and at most play a role as a second opinion. Complete integrated dermatological image analysis systems are rarely if ever found in clinical use [5, 7].

1.1 Key Limitations

Currently, features based on colour and texture associated with colour play a dominant role in automatic CSLD [2]. This is on one hand because colour is an important diagnostic indicator. As stated in any clinical dermatology textbook, colours seen on the skin surface reflect many aspects of its internal structure and composition, such as the amount of epidermal melanin, dermal blood and concentration of melanocytes invading the papillary dermis [8, 9]. On the other hand, this is also because of the limitation of imaging techniques since feature extraction is highly reliant on the information captured in lesion images. From reviewing the relevant literature, it can be seen that most computer-aided diagnosis systems are developed using dermoscopy (Epiluminescence microscopy) images (detailed in Section 2.2.1(2)). The advantage of dermoscopic images is that the skin subsurface structure is magnified so that more details are visible [6, 10]. The disadvantage is that the oil immersion process and the pressure applied on the skin at the dermoscopy interface can distort the elastic skin structures[11], *e.g.,* the applied oil may smooth the rough appearance of skin surface. Also, the applied liquid sometimes adds artifact (*e.g.,* air bulbs) to lesion surface. These would degrade the image quality of morphological features. Most importantly, this pin-hole camera based imaging system transforms the 3D world-space to the 2D imaging-plane. During this transformation, useful information may be lost, *e.g.,* the lesion superficial shape information. Comparatively, range sensors such as stereo-photometric (see Section 2.2.1(5)) have the advantage of delivering such information by preserving it in the associated 3D or depth data. The 3D (or depth) information could be vital for certain tasks since there is no doubt that the process for depth recovery is a part of the human visual system in reality. In recent decades, an especially desirable capability in computer vision is the automatic reconstruction and analysis of the surrounding 3D environment and recognition of objects in that space [12].

Besides, skin is the outermost tissue of the human body whose surface is characterised by polyhydric mesh structures representing the three dimensional organization of the dermis and the subcutaneous tissue [13, 14]. The topography of the skin is directly related to the cell growth patterns under the skin surface [11]. When the skin is healthy, this topical structure is highly regular. When skin problems arise, it becomes irregular. Taking malignant melanoma for instance, the growth of abnormal cells in the upper dermis will result in irregular clumps that disrupt the regular skin pattern on the surface. Hence, observing the changing of surface pattern would help to locate the problem skin region (or lesion). On the other hand, the different pathogenies of different skin problems, such as the cell of origin usually result in different topographical appearances of the outermost skin surface. For example, the common melanocytic nevi appears different from melanoma and BCC with an unruffled, smooth landscape and a circular or oval shape. In contrast, melanoma often presents the geographical appearance of rough crests, canyons and reefs [15]; while BCC has a persistent, non-healing and eroded area with poorly defined borders. The topography of the skin surface is considered as a mirror of the functional skin status [16]. It can be deemed as another important skin descriptor similar to colour, revealing delicate differences within lesions and playing an important role in dermatological diagnosis, in terms of lesion localization and classification. In addition, as an intrinsic factor that concerns the physical properties of an object, the 3D shape has a particular advantage of being free from the influence of the external environment, such as illumination conditions [17]. This is a desired characteristic of features used for object recognition. As a result, we hypothesize that ***the surface shape embodies important information that serves to represent lesion identity and status. Combining the surface shape and colour based features would improve the performances of CSLD systems on the lesion segmentation and classification.***

In fact, examining the surface shape appearance is also a part of the clinical lesion diagnosis and it is done by touching and observing from different points of view.

Since the computer-based diagnostic systems play a role as a 'clinical eye' that mimics and augments the clinician's ability to distinguish between lesions using various computer vision techniques, we argue that it would be incomplete without taking into account the lesion surface shape information. We believe that through capturing new and better information, a 3D based system may open new possibilities in improving the medical analysis of skin lesions, although there have always existed arguments regarding whether depth is necessary in an object recognition, or if it is only of limited use when it is available. For the skin lesion diagnosis application, even the dermatologists could not give a decisive answer of the importance of depth information during lesion diagnosis [1]. The ultimate answer to whether to use depth in a system can only be determined by concrete experiments [18].

The aim of the thesis is therefore to answer the above concerns and to testify the above claim through extensive experiments on the lesion datasets (see Appendix D) captured using the Dense Stereo Imaging System (DSIS, detailed in Section 2.2.1(5)) developed by Dimensional Imaging [19]. Specifically, we need to investigate whether the extracted skin shape properties would be of potential benefit in the segmentation and classification process with the addition of 3D data. We have not identified any previous non-Edinburgh research in the literature that have explored this specific question, however, there has been some previous research using 3D, *e.g.,* the UWE's group represented the lesion line pattern property using the surface norm data [11] (as discussed in Section 2.3). Besides, unlike most studies in the literature that only take into account the pigmented lesions and solve the binary classification problems [20, 21], *i.e.,* distinguishing melanoma from melanocytic nevus, in our research, multiple pigmented and non-pigmented lesion classes are included in our databases. They are Actinic keratosis (AK), Basal cell carcinoma (BCC), Squamous cell carcinoma (SCC), Melanocytic nevus (ML) and Seborrheic keratosis (SK). In addition to two skin cancers, three benign lesions are also taken into account. Melanocytic nevus is commonly known as mole. It requires a dermatologist to fully evaluate it because it is often mistaken for melanoma. AKs are a focal area of dysplasia within the epidermis. They are not malignant but occasionally develop into SCC. SKs are benign clonal tumours of the skin. Although they may be of cosmetic importance, they are of no medical importance except that they are frequently confused with other tumours. Considering multiple lesion classes undoubtedly increases the diagnosis difficulty. From this respective, including

[1] Private discussion with Dr. Jonathan L Rees, the Grant Chair of Dermatology at the University of Edinburgh

extra diagnostic information appears more important. The melanoma is excluded in the database, because 1) it has been well studied in many other works and 2) there are very few samples collected at the Dermatology Department of Edinburgh University using the 3D capturing system over the past four years. Though most of the lesions (two types of skin cancer and three types of benign lesion) in our database are not as deadly as melanoma, their frequent appearance in the clinic still raises concerns. In addition, compared to the intensive studies on melanoma, the analysis of these skin conditions are fairly rare.

1.2 Thesis Overview

1. In order to validate the importance of depth information to lesion identification (or segmentation), we incorporate diverse image cues (*i.e.,* colour, depth and texture) into a segmentation model to investigate whether or not the extra depth information would lead to better results. A region-based probabilistic formulation of the deformable model that is implemented within the level-set framework is built as a testing platform. The experiments on 50 skin lesion belonging to five lesion classes (AK, BCC, SCC, ML and SK) show that integrating depth based properties results in an overall improvement of segmentation, in terms of both accuracy and consistency. The error rate is reduced from $7.80\% \pm 5.35\%$ to $6.62\% \pm 2.60\%$. The result reveals that depth is complementary to colour and it does improves the lesion identification, particularly for non-pigmented lesions which have less colour variation over different regions. In order to optimize the usage of different feature combinations for segmenting different types of lesions, we propose a hierarchical strategy which further improves the segmentation performance. Parts of the results have been published in [22]. More details can be found in Chapter 4.

2. The error rate given above is obtained by comparing the computer-based segmentation result against a reference or Ground Truth (GT). In general, the GT refers to the manual segmentation of a clinical expert. To account for individual subjectivity, usually the GT is estimated from multiple manual segmentations. In order to take into account the inter-rater variation, we propose a novel GT generation algorithm that maximizes the *a posteriori* probability and incorporates the segmentation pattern information (LSMLP). This approach integrates

the prior analysis result of human segmentation patterns and solves the problem in a form of energy optimization within a level-set framework. The experiments on both synthetic and real data show that this approach has larger chance to find an accurate ground truth. The corresponding content can be found in Chapter 3. Some results have been published in [23, 24].

3. For the purpose of evaluating the contribution of 3D information on lesion diagnosis, we carry out extensive and rigorous tests on both human and computer based classifications. The experiments comparing human performance support our claim as the diagnosis results show a significant increment of 8.5% when using the 3D images. We further investigate the contribution of 3D data to a computer based five non-melanoma skin lesion classification task by a comparison between using colour features only and using both colour and depth features. The comparison results on six classifier models and two databases all show that the addition of depth does improve the diagnostic accuracy. Based on an experiment using a Support Vector Machine (SVM) on DATABASE II (see Appendix D), adding the 3D-based features gives an improved classification rate of 80.67% compared to only using colour features (75.25%). This improvement is also proved to be statistically significant. Details can be found in Chapter 5.

Chapter 2

Literature Review

This chapter presents an overview of existing work in the skin lesion diagnosis field. Firstly, Section 2.1 provides a brief review of the computer-based lesion diagnosis development; Section 2.2 then explores the state of the art techniques and systems in the field including: 1) the imaging techniques that have been employed in computer-based diagnosis systems to capture various kinds of lesion information, 2) the popular diagnostic schemes that have been proposed to distinguish different skin conditions as well as the features that are used to represent different criteria and 3) the existing computer-based skin lesion diagnostic systems. The final section of this chapter reviews the studies that apply 3D data to analyse skin lesions (Section 2.3).

2.1 History

In the clinical environment, there are often recognizable precursor conditions for different lesions. By examining these conditions through inspection and palpation, dermatologists are able to make a presumptive prognosis of a lesion. However, the diagnosis given by clinical experts has the disadvantages of being subjective, descriptive and expensive. For the purpose of achieving an objective, consistent, quantitative and cheap diagnosis, many Computer based Skin Lesion Diagnosis (CSLD) systems have been developed. The first related work dates back to 1987 [4], in which the authors suggested the computer could be a possible new tool for analysing melanoma, regarding its two advantages: 1) allowing a standardized and repeatable and objective analysis of lesion images and 2) the capability of analysing details not perceivable by the human eye. Several criteria extracted using classic digital image analysis techniques were applied to analyse a few melanoma cases and they were shown to add valuable information to the diagnosis. This preliminary study stimulated the continuation of research since then. In 1991, Green *et al.* [25] presented a pioneering fully automated melanoma diagnosis system [20]. The system analysed 70 pigmented lesions captured by a colour video camera in terms of eleven clinical and histological characteristic correlated measurements, based on which the lesions were assigned to different classes through a discriminant analysis. Though the performance of the system was hindered by the simple and crude computer vision and machine learning techniques as well as the imaging quality, this study generated a prototype of the pigmented lesion diagnosis system that comprised three major steps: 1) determination of lesion region, 2) extraction of diagnostic features and 3) building classification models with important features. Each step itself has formed an interesting and challenging research topic and has been widely discussed. Detailed reviews of these topics are in the literature review section of the respective chapters (3, 4 and 5).

The turning point in the development of CSLD systems was the emergence of the dermoscopy technique in 1994. Day *et al.* [6] categorized the developments of CSLD into two periods, pre-1995 and post-1994 according to this factor. Prior to 1995, conventional naked-eye images were used as the major input for CSLD. From 1995, CSLD systems commenced to use the dermoscopic-based image set as their input [5]. As dermoscopy permits the visualization of new and better morphological features which in most cases facilitate early diagnosis, it boosts the performance of CSLD systems and has been considered as the state of the art lesion data capturing technique. A systematic review covering Medline entries from 1983 to 1997 revealed that dermoscopy had $10-27\%$ increase in sensitivity [20].

2.2 The State of The Art

2.2.1 The Imaging Techniques

1. **Conventional macroscopic camera** The use of commercially available photographic cameras is still quite common in skin lesion inspection systems, particularly for telemedicine purposes [2]. For example, a recent study of the macroscopic skin feature - skin pattern was carried out by She *et al.* [26] based on the simply captured white light optical clinical skin images. The researchers measured the skin pattern disruptions in terms of the local line direction using the local isotropy metric. Their work indicated that combining the local line

direction descriptors and the ABCD features might be a promising method to distinguish malignant melanoma from benign lesions [26].

The problems of using conventional images are the poor resolution for small skin lesions and the difficulty in handling the environmental variation influences (*e.g.,* illumination). As a result, newer technologies that could solve these problems and provide new and better information come to the forefront in providing greater diagnostic accuracy.

2. **Dermoscopy** Dermoscopy is a common tool that is used in the clinical examination of pigmented skin lesions, *i.e.,* distinguishing melanomas from benign melanocytic nevus (moles) and seborrheic keratosis (Seborrheic keratosis are actually benign keratinocyte tumours which have normal numbers of melanocytes, but which are hyperpigmented due to over production of melanin in the normal number of melanocytes there). It is a non-invasive skin imaging technique that uses optical magnification and either liquid immersion and low angle-of-incidence lighting or cross-polarized lighting to make the contact area translucent and make subsurface structures (*e.g.,* dermal features) become visible compared to conventional macroscopic (clinical) images [20, 2]. Its advantages are the capabilities of 1) providing a more detailed inspection of the surface of pigmented skin lesions and renders the epidermis translucent, making many dermal features become visible [2] and 2) avoiding the defused reflection on the skin surface thanks to the oil immersion process. However, it has also been criticized for 1) being difficult to learn and subjective, particularly in the hands of inexperienced dermatologists and 2) easily distorting the elastic skin structures and degrading the image quality of morphological features because of the pressure applied on the skin at the dermoscopy interface and the oil immersion [11]. Several studies have confirmed the limits of unaided dermoscopy [27], however, its good imaging properties have made it the most popularly used in computer aided pigmented lesion analysis systems [28, 20, 29, 30, 31, 32, 21, 33]. MoleMax is the most frequently used digital dermatoscopy system. The patented light polarization technique enables the dermatoscopic images to be captured without the use of immersion fluids. In addition, the specially developed system camera in MoleMax allows other capturing modes, like macro or even close-up imaging [34].

   Recently, some studies have focused on the enhancement of dermoscopy images as a pre-processing step in CSLD systems. For instance, in order to obtain better features for skin lesion segmentation, Schaefer *et al.* [35] introduced an algorithm that could enhance the colour information and image contrast. Their method solved the problems of poor contrast and lack of colour calibration which were often encountered when analysing lesion images and improved the segmentation performance which is a critical step in analysing skin lesion images. In order to recover the colour information from the inappropriate white balance or brightness in the image, Iyatomi *et al.* [36] proposed a supervised learning approach to calibrate the colour of a new dermoscopy image based on the Hue-Saturation-Value colour model. The modified colour distribution of a given image was closer to that of the training image set.

3. **Spectrophotometric Intracutaneous Analysis (SIAscopy)** SIAscopy was introduced by the Medical Image Analysis Group of Birmingham University as a new technique for imaging pigmented skin lesions [8]. This optical skin imaging method acted as a non-invasive biopsy and allowed an insight into the skin histology *in vivo*. SIAscopy produced eight narrow-band spectrally filtered images of the local skin region with radiation ranging from 400 to 1000*nm*. They related to the skin internal information regarding total melanin content of the epidermis and papillary dermis and collagen, *etc*. New features extracted to represent this information were found to be highly specific (80.1%) and sensitive (82.7%) for melanoma in a database of 348 pigmented lesions (52 melanomas). The authors concluded that SIAscopy delivered new and useful information to the diagnosis of pigmented skin lesions. In addition, experiments also showed that SIAscopy compared very favourably with dermoscopy when analysed using receiver-operator characteristic curves.

4. **Photometric Stereo Device (PSD)** Researchers in the Machine Vision Laboratory, University of the West of England employed a six-light photometric stereo device to analyse malignant melanoma. PSD was equipped with a camera and six LED light sources. It captured six separate images with each LED independently illuminated. The system output three 3D surface normal images and the surface reflectance map (see *Fig.* 2.1) [11]. The surface normal images enabled the analysis of skin surface textures. Some studies based on PSD are reviewed in Section 2.3.

5. **Dense Stereo Imaging System (DSIS)** Our group (Machine Vision Unit, School

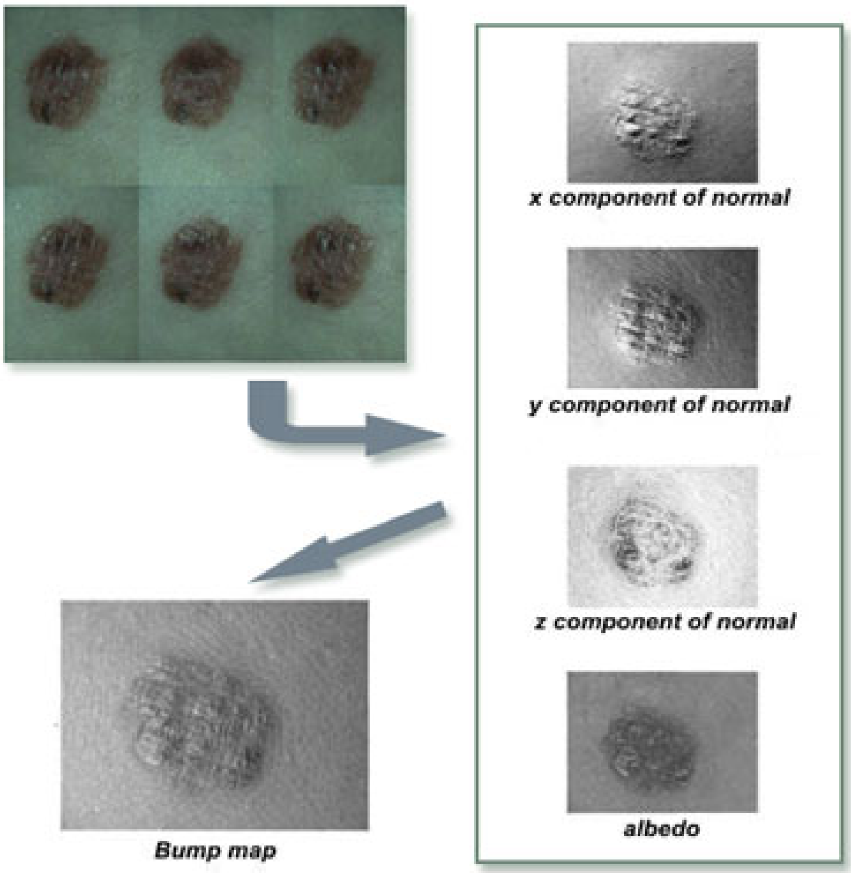


Figure 2.1: An example of the captured image set using the six-light photometric stereo device from [37]

of Informatics, Edinburgh University) also employs 3D imaging equipment, the Dense Stereo Imaging System (DSIS) (shown in *Fig.* 2.2) developed by Dimensional Imaging [19] to capture the complete 3D data of skin lesion surfaces. DSIS acquires stereo-pair images of the lesion and then decodes the depth information explicitly captured within the stereo-pair. The colour image of lesions can also be captured simultaneously and is 1:1 aligned with the 3D image (see *Fig.* 2.3). Aiming at measuring the micro structure of the skin lesion (the size of skin lesion varies and typical lesion size is larger than $10mm \times 10mm$), the system is configured for a small area of $80\ mm \times 60\ mm$ and set to a fixed focal distance. Raw resolution is about 30 μm in x, y and z directions. However, as a consequence of image smoothing, the system accuracy test presented in Appendix B shows that DSIS is able to **detect** and **separate** fine textures with $0.7mm$ scale. In contrast to PSD, DSIS obtains complete 3D data, with which one could extract various kinds of surface morphological descriptors, including the surface texture based properties for distinguishing skin lesions.

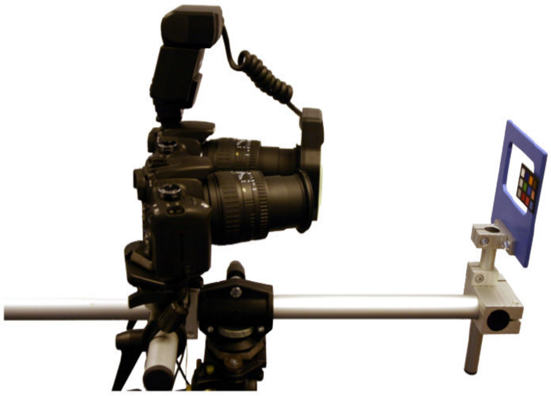


Figure 2.2: The dense stereo imaging system (DSIS)

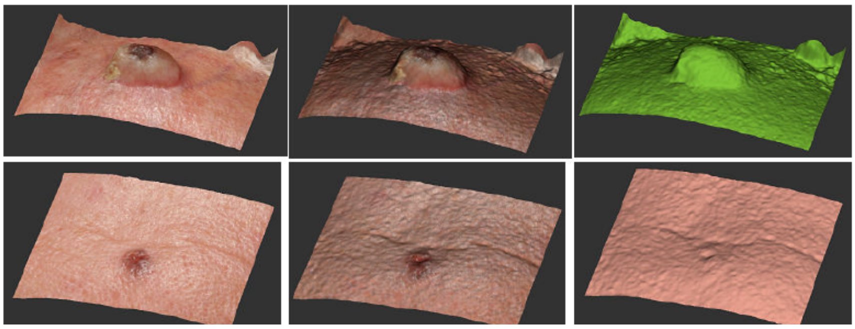


Figure 2.3: The 3D data of skin lesions SCC (Top) and BCC (Bottom) reconstructed from the 3D data captured using DSIS. From the left to the right, they are textured 3D model, textured 3D model with lighting and non-textured 3D model with lighting

For a review of other imaging systems (*e.g.,* Video RGB camera, multi-frequency electrical and raman spectra), we refer the reader to [2].

2.2.2 Diagnostic Schemes

Skin conditions often have certain distinctive features that, in most cases, will enable the doctor to recognise the disease. These practical criteria are organized into several schemes, based on which the computer-based diagnostic properties are derived.

1. **ABCD Rule** The ABCD rule introduced by Stolz *et al.* [33] has been accepted as a standard for distinguishing melanomas from melanocytic nevus with dermoscopy. The ABCD acronym stands for Asymmetry (meaning one half of the mole is different from the other), Border irregularity (the edges or borders of melanomas are usually ragged or notched), Colour (melanoma often has a variety of hues and colors within the same lesion) and Diameter (most melanomas are usually greater than 6 mm in diameter when diagnosed, although they can be smaller). This rule was designed more than 20 years ago specifically for early melanoma diagnosis. It offers a standard that aids in distinguishing potentially cancerous pigmented lesions (melanoma) from benign pigmented moles (ML). To date, the majority of features used in CSLD systems are defined to represent the four criteria of the ABCD rule. For instance, the asymmetry parameter (A) was modeled by two Symmetry Distances (SD), the basic SD and the fuzzy SD, as well as the simple circularity of the shape of a lesion in [38]. Grana *et al.* [39] proposed mathematical descriptors like lesion slope, lesion slope regularity, *etc.*, to measure the skin-lesion gradient. The efficacy of these border features were assessed through a classification of a database containing 510 pigmented skin lesions (85 melanomas and 425 nevus). The result showed that these features helped to achieve a sensitivity of 85.9% and a specificity of 74.1%. Researchers in the Medical Image Analysis Group, Birmingham University proposed the Irregularity Index as a measure of border irregularity that was considered as an significant diagnostic factor when assessing a lesion for malignancy [40]. The extracted features were applied to different between the melanoma group and the benign lesion group using a linear classifier and achieved a classification rate of 82.4%. To account for the colour variation (C), the authors in [41] proposed 15 significant colour based descriptors that were obtained by the median cut colour quantization method. Stanley *et al.* [42] proposed percent melanoma colour and colour clustering ratio features using a colour histogram analysis technique. Manousaki *et al.* [15] introduced colour textural roughness based features - fractal dimension (a measure of the irregularity of a given surface) and lacunarity (inhomogeneity within a fractal surface).

   Even though most of the studies claimed the effectiveness of their proposed ABCD based features, it is not clear yet which features are more informative according to a survey carried out by Maglogiannis *et al.* [2], in which no agreement could be made. Moreover, the definitions of some features are inconsistent and there is a lack of rationale for most given feature functions [10]. On the other hand, this subjective ABCD rule itself has not yet achieved high consensus among expert dermatologists, as not all melanomas follow the ABCD rule and some other lesions might fit in the ABCD rule. For example, the Nodular Melanoma (NM), which commonly occurs as symmetric, elevated lesions that are uniform in colour and non-pigmented does not fit the ABCD criteria for melanoma diagnosis [43]. As a result, other diagnostic criteria taking into account more morphologic parameters are needed, *e.g.,* the 7-point checklist [44].

2. **7-point Checklist** The 7-point checklist scores a lesion's malignancy (*i.e.,* being melanoma) using standard dermoscopic criteria, including the atypical pigment network, gray-blue areas, atypical vascular pattern, radial streaming or streaks, irregular dots and globules, regression pattern and irregular diffuse pigmentation. This criterion reveals a possibility of lesion diagnosis based on the recognition of certain morphological patterns. The key to use this criterion is to first detect these patterns. For example, Betta *et al.*[45] showed an example for estimating the pigment network and atypical vascular pattern of a lesion in. In [46], the authors proposed a machine learning approach to the detection of blue-white veil structures (irregular, structureless areas of confluent blue pigmentation with an overlying white 'ground-glass' film) in dermoscopy images, given the blue-white veil is an indicator of the melanoma. In [21], the authors proposed a semi-automated melanoma identification method that was based on the finding that granularity was significantly associated with melanoma. However, this approach needed human involvement in marking the granularity or regions as close as possible to granular spots in melanoma and non-melanoma lesions, respectively. As a result, this method has the disadvantages of subjectiveness and being labour consuming. Dalal *et al.* [28] achieved an automatic melanoma discrimination from benign lesions (ML) with the assistance of white areas. By extracting features from the detected white areas and putting them into a back-propagation neural network, they achieved 95% diagnostic accuracy using the data set including 57 melanomas and 187 benign nevi. These dermoscopic feature based methods are comparatively new in CSLD. As the prerequisite of using this strategy is the detection of these structures, which is not a trivial task, it has not been as commonly used as the ABCD rule. But once the techniques for the detection of these patterns mature, it is very likely that the diagnostic performance would

be boosted.

The above two diagnostic schemes are the state of the art standards for the diagnosis of melanoma. Their common property is being limited to the diagnosis of pigmented lesions, particularly in distinguishing between melanoma and nevi. In addition, they are dermoscopic schemes. Hence, features extracted based on ABCD rule and 7-point checklist are limited to the diagnosis of pigmented lesions and are mostly associated with colour information in various colour spaces[2]. To further improve the performance and generalize the diagnosis ability (*e.g.,* diagnosing non-pigmented lesions) of CSLD, new diagnostic criteria need to be included. For example, several studies realized that the disruption of skin pattern would be an important indicator of lesion malignancy. She *et al.* [26] explored this factor by introducing local line direction descriptors based on the local isotropy metric. Zhou *et al.* [47] characterized the disruption using 3D differential forms. Both projects concluded that this new research direction, which was orthogonal to the ABCD rules, would add information useful for the diagnosis of melanoma.

2.2.3 Computer Based Skin Lesion Diagnostic Systems

In recent years, with the application of dermoscopy and the rapid development of the medical image analysis techniques, many CSLD systems have been developed. *Table.* 2.1 lists several recent studies in automated melanoma diagnosis. Menzies *et al.* [48] evaluated the performance of an automated dermoscopy image analysis instrument, SolarScan, for the diagnosis of primary melanoma. Based on the diagnosis results on a test set including 78 lesions (13 melanomas), SolarScan gave a sensitivity (SE) of 91% and specificity (SP) of 68% for melanoma and it had comparable or superior performance when compared with clinicians. The CSLD system presented by Iyatomi *et al.* [49] was directed at non-white populations for acral volar melanoma detection. The reported evaluation accuracy shows a SE of 81.1% and SP of 92.1% (considering unsuccessful lesion segmentation cases as false classification; otherwise, a SE of 100%, SP of 95.9%). The automatic melanoma diagnosis system proposed by Celebi *et al.* [10] yielded a Specificity (SP) of 92.34% and a Sensitivity (SE) of 93.33% based on a set of 564 dermoscopic images. In [28], the author proposed a melanoma identification algorithm based on the dermoscopic feature of white areas and reported a diagnostic accuracy of 95% using the data set including 57 melanomas and 187 benign nevi. To open the computer based lesion diagnostic resources to the public, Iyatomi *et al.* [20, 30] introduced an internet-based melanoma screening systems. In order to obtain a diagnostic result, the user could upload a dermoscopic image, together with the associated clinical data, *e.g.,* lesion position. The system would extract the lesion area, calculate the lesion characteristics and yield a diagnosis in the form of a malignancy score between 0 and 100 based on the output of a linear classifier or artificial neural network. The latest version of their system [20] featured a sophisticated dermatologist-like lesion segmentation algorithm that attained superior performance. In addition, the system could automatically selected the appropriate diagnostic classifier from linear and back-propagation artificial neural network based on the location of lesions provided by users. Based on a leave-one-out cross-validation test on a set of 1258 dermoscopy images (1060 nevi and 198 melanomas) using 428 image related objective features categorized into asymmetry, border, colour and texture properties, the system achieved 85.9% sensitivity, 86.0% specificity. In our group, Ballerini *et al.* [50] conducted research on a Content-Based Image Retrieval system (CBIR), with the aim of providing a diagnostic aid for skin lesion prognosis. The decision was made by humans, but supported by retrieving and displaying relevant past cases visually similar to the one under examination. More information about CBIR systems can be found in [29].

From *Table.* 2.1, it can be seen that some CSLD systems reported higher diagnostic performances (with sensitivity and specificity of about 90%) than clinical experts ($75-84\%$ in diagnosing melanoma [20]). However, these systems still have several limitations regarding the acceptable lesion cases (most systems can only analyse pigmented lesions and distinguish between melanoma and melanocytic nevi) and using limited diagnostic information (dermoscopic-images based). The diagnostic capability of current automated systems does not yet match that of an expert dermatologist [20].

2.3 Skin Lesion Studies using 3D Data

Recently, with rapidly growing research in 3D computer vision, the analysis of objects in 3D space becomes possible. Some studies that apply 3D systems to skin analysis problem are:

1) Castellini *et al.* [53] is possibly the first group carrying out real 3D measurement of the superficial structure of skin lesions. As the second growth phase of the melanoma, vertical growth is considered to be important clinical prognostic information. In the light of this, the authors were inspired to assess the lesion height measured using a

Source	Segmentation method	Classifier	Images #	Classes	Performance	Image source

Green *et al.* [25]
1991
	-	Discriminant analysis	70	melanoma(5)
				nevus(53)
				other pigmented lesions (12)
					DA=76%	video camera

Rubegni *et al.* [51]
2002
	-	ANN	588	melanoma(217)
				nevus(371)
					DA=94%	Dermoscopy

Menzies *et al.* [48]
2005
	-	Logistic	78	melanoma(13)
				nevus(65)
					SE=91%
					SP=68%
						Dermoscopy

Celebi *et al.* [10]
2007
	region growing	SVM	564	melanoma(88)
				nevus(476)
					SE=93.33%
					SP=92.34%
						Dermoscopy

Stanley *et al.* [42] 2007	manually drawn	threshold-based discrimination	226	melanoma (113) nevus(113)	SE=87.7% SP=74.9%	Dermoscopy
Iyatomi *et al.* [49] 2008	region growing	linear classifier	213	acral volar melanoma(37) nevus(176)	SE=81.1% SP=92.1%	Dermoscopy

Iyatomi *et al.* [30]
2008
	dermotolgist-like
	region growing based
		ANN	1258	melanoma(198)
				nevus(1060)
					SE=85.9%
					SP=86.0%
						Dermoscopy

Iyatomi *et al.* [52] 2010	threshold+ morphological operations	linear classifier	655	melanocytic lesions (548) non melanocytic lesions(107)	SE=98.0% SP=86.6%	Dermoscopy

Table 2.1: Recent studies on computer aided skin lesion diagnosis

laser triangulation technique, which had the disadvantage of a long capturing time. Even though their 3D measurement system enhanced the knowledge in the field of measurement and reconstruction of skin characteristics, their work mainly focused on proving the ability of the system to capture the morphological characteristics of the lesion. There was no further discussion about the diagnostic value of the shape of the lesion.

2) Leveque *et al.* [54] introduced a new integrated tool DERMA that allowed the measurement of chronic wounds based on the 3D data obtained using a laser triangulation scanner. DERMA was successfully used to monitor the development of lesions, *e.g.,* the wound healing process.

3) In [55], the authors described a new method (SkinChip) that used a non-invasive 3D-based device to characterize the properties of the skin surface. As the captured skin surface data allowed a highly precise observation of the skin topography that could be easily quantified in terms of line density and line orientation, SkinChip was considered as a convenient way to evaluate the age wrinkle smoothing with regard to hydration. This study showed the potential benefit of 3D data on skin analysis.

4) Ding *et al.* [11] applied the Photometric Stereo Device (PSD) to capture the skin surface data in the form of surface normal images, which were then used to analyse disruptions in skin patterns that were found to be larger for melanoma than for benign lesions. In contrast to analysing the skin line patterns on 2D images as in [26], their work was based on 3D skin surface normal information. The 3D skin pattern disruption related features were extracted as the residuals between the surface normal

data and those from a best-fit model. Their experiment on a database including 12 melanoma and 34 benign lesions showed that using 3D skin surface normal features improved the specificity from 25.7% to 91.7% compared to using 2D skin line pattern features. The sensitivity did not change. The results demonstrated that the addition of 3D normal features lowered the risk of considering a benign lesion as melanoma and indicated the non-invasive 3D test could improve the accuracy of melanoma diagnosis. Another study from their group [47] investigated the effectiveness of four 3D-based curvature pattern related properties in melanoma diagnosis. A test on a small-scale data set comprised of 23 melanoma and 53 benign lesions indicated the effectiveness of the 3D curvature pattern in melanoma diagnosis, though the improvement was without sufficient statistical proof when compared to the classic 2D features.

5) The most closely related work to our research is the preliminary study carried out in our lab by McDonagh *et al.* [56]. In their study, the accuracy of the diagnosis obtained by combining the topography and colour features were compared with that obtained by using only colour features. The comparison was based on the Bayesian classifier with a unimodal multidimensional Gaussian class model. The feature set was selected using a greedy forward selection strategy that started with an empty set of features and progressively added features. Their experimental results suggested that the depth information might improve the diagnostic rate. Though a) this conclusion was weak regarding the statistical significance (with a p-value of 0.3), b) the experiment was based on limited number of lesions (234 sample in total) and c) the ill-posed problem of inverting the covariance matrix in the Bayesian classifier limited the number of features that could be used in the classification. But still, this work for the first time indicated that the effective 3D-based information could open new possibilities in improving lesion diagnostic rate.

2.4 Summary

In summary, one can find that

1) much research on automated melanoma analysis has been done and has achieved acceptable performance levels. However, that research was only designed to distinguish between melanoma and melanocytic nevi and was based on the analysis of chromatic information (mostly derived from dermoscopic images).

2) there has emerged some skin lesion analysis using 3D data. Recently, the idea of including 3D based properties into the lesion diagnosis has perked up. But no concrete

and systematic studies showing the benefit of 3D information could be found.

Therefore, applying 3D data and 3D-based analysis methods to non-melanoma skin cancers or other skin conditions is an open research area. The development of new 3D imaging systems which could capture good surface shape and colour information simultaneously allows us to explore this research topic in this thesis.

Chapter 3

Ground Truth for Segmentation Evaluation

Having ground truth is critical for evaluating segmentation algorithms and estimating the ground truth from a collection of manual segmentations remains an open problem. This chapter first presents an overview of the existing ground truth estimation approaches in Section 3.2, followed by an analysis of manual segmentations which aims at obtaining a better understanding of the pattern of the inter-rater variation in Section 3.3. Section 3.4 then proposed three ground truth estimation approaches based on different energy function formulations and solved by optimization under a level-set framework. Experiments on both synthetic and real data presented later, in Section 3.5, show that the approached methods are promising in finding an accurate ground truth, particularly for the one that integrates the prior shape analysis result.

3.1 Introduction

Segmentation is the first step of the computer-based skin lesion diagnosis algorithms and its accuracy is of crucial importance for the subsequent analysis. Numerous computer-based skin lesion segmentation methods have been developed based on different methodologies. In order to analyse their performance, objective evaluation methods are needed.

In general, the segmentation evaluation can be categorized into two groups: supervised and unsupervised evaluation, depending on whether the method utilizes *a priori* knowledge, which we refer to as the Ground Truth (**GT**). The former is more widely used in the medical field. It considers the accuracy of the segmentation result as the degree to which the result corresponds to the ground truth segmentation through evaluation metrics. Unfortunately, the **GT** normally does not exist in practice, in the absence of which, the segmentation evaluation is impossible. Although synthetic data or phantoms [57] help, they do not allow the reproduction of the full range of characteristics observed in clinical data and lose fidelity [58]. An alternative way is to compare the method output against segmentation made by a trained rater. Because a single expert, even experienced, is likely to be subject to a personal bias and poor precision, therefore cannot be used as an absolute reference. The comparative ground truth must be a good compromise within a group of raters [59].

However, the inter-rater segmentations show a significant disagreement according to the rater's subjective criteria in placing the boundary [59, 24]. Hence, the question is raised as *how to compensate for this inter-rater variability, and derive a ground truth from the results given by multi-raters*. To date, the most appropriate strategy to combine such segmentations is unclear and it has become a popular research topic itself [1].

3.2 Literature Review

Drawing on the concept from the field of pattern recognition, deriving the ground truth from a collection of manual labeling results is a decision fusion problem. Treating the labeling results of a target image as classifications, a combinational modal value for each pixel can be obtained using a decision fusion rule. An appropriate fusion rule should be able to compensate for the inter-rater differences and eliminate the intra-rater variability, which are also referred to as the systematic and random errors in the individual results. Therefore, the Estimated Ground Truth (**EGT**) arrives at a consensus that is closer to the truth than any of the constituent segmentations.

As a means of decision fusion, the Voting-Rule-based strategy is the most straightforward. For the category labeling representation, such as binary labeling, the *Majority Voting Rule* (**MV**) assigns each pixel to the value that the majority of experts agree upon. This method is the most popular in the literature and it has the advantages of being simple and efficient. Unfortunately, the **MV** is directly related to human intuition and does not provide guidance as to how many experts should agree before making the decision [1]. If a continuous label assignment like probabilities is supplied, more complex voting rules, such as Product Rule, Sum Rule, Max Rule, *etc.,* can be applied [60]. However, all these strategies treat each rater equally and do not allow the incorporation of *a priori* information that describes the performance of individual raters. Though a weighted version of the vote-rule strategy can be easily implemented, a robust way to derive weight parameters has not been clarified.

Warfield *et al.* are the pioneers in estimating the **GT** incorporating the rater's performance as weights. Their **STAPLE** algorithm [1] is so far one of the most referenced approaches in the field. **STAPLE** treated decision fusion as a maximum-likelihood problem. It was solved using the Expectation-Maximization(EM) algorithm that guaranteed convergence, but not necessarily global optimality. **STAPLE** gave the quantitive estimate of the performance level parameters of raters in terms of the sensitivity and specificity and, based on which, it could output a probabilistic estimation of the **GT** simultaneously. Commowick *et al.* [61] further investigated the factors that might influence **STAPLE** performance. They reported that the initialization on the performance level parameters affected the estimated results heavily and they also estimated the confidence interval of the estimated performance parameters. Furthermore, the experiments revealed a dependence of the confidence intervals with respect to the number of voxels and the size of the segmentation structure of interest. According to their conclusion, a sufficient number of manual results should be used for **STAPLE** to produce precise results. In their particular case, the uncertainty of performance parameters was stable when 5 or more experts were used in the study. Warfield *et al.* [58] generalized **STAPLE** by extending it to the cases where the boundary could be represented by a signed distance transform or level-set where each pixel has a continuous score. Their formulation considered the subjective bias in terms of overestimate or underestimate of the position of a boundary as well the consistency of the raters' performance level parameters. Langerak *et al.* [62] and Klein *et al.* [63] highlighted that the performance of **STAPLE** was application dependent. It failed when the performances of the raters varied greatly. This can be explained by the fact that, even though fusing results in a weighted way, **STAPLE** takes into account all raters. A bad rater can contaminate the overall result, especially when an inappropriate initialization is allocated. In this context, adding a rater selection step helps. In [62], the authors proposed a simplified **STAPLE** variant. This variant iteratively selected the optimal segmentation results based on image similarity measures and abandoned the ones with poor quality due to the wrong registration result. The final result was a combination of the optimal segmentations in a weighted Majority Vote procedure. The selection step helped to deal with the large variability problem encountered in their application and hence produced better results than **STAPLE**. Their approach required a large number of manual results because of their abandonment strategy and several parameters needed to be tuned in the iteration step, like the number of segmentations to be discarded and the similarity threshold. Moreover, the algorithm had no guarantee of convergence. It does however give us a hint as to whether or not a prior study of the segmentations would help.

From our point of view, the key difference to the decision fusion function is how the weight that reflects the rater's performance is set up. If the weights are uniform, it is assumed that the performance level of each rater is identical. The **EGT** should be a compromise between the raters which minimizes the overall discrepancy between the ground truth and individual segmentations measured using a certain metric. Various metrics exist in the literature and can be categorized into two groups: 1) spatial-overlap-based, which measures the overlap between two segmentations, *e.g., XOR* [64], Dice Similarity Coefficient (*DSC*) [63], Overlap Ratio (*OR*) [65] and 2) boundary-difference-based, *e.g., FOM* [66] and Hausdorff measure, which accounts for the distance between two boundaries. These metrics are also useful in comparing the computer-based segmentation against the estimated ground truth.

If it is assumed that the performance level of raters varies and should be accounted for in the estimation, then the weights are set differently. Large weights are assigned to high quality raters and low weights to poor quality raters. The difficulty lies in *how to characterize the performance level and how to embed them into a **GT** estimation model?* As mentioned previously, the segmentation by the human rater is subject to both systematic and random errors in practice. The latter results in the intra-rater variability in the placement of individual labels [65] and it is defined as the precision in [58] measuring the reproducability of individual raters. It can be diminished when multiple segmentations are fused [65]. The systematic error arises during the stage of defining the reference boundary and leads to the discrepancy between the reference labeling and the unknowable true labeling. It reflects a consistent bias over the position of a boundary of a rater [58]. This error totally depends on the rater's subjective segmentation policy. Previous algorithms like **STAPLE** intend to characterize such a bias through performance level estimation in terms of sensitivity and specificity parameters and compensate for it in the weight-based fusion process. However, we hypothesize that the estimated ground truth cannot arrive at the true position unless the following condition is satisfied: both overestimation and underestimation exist and are equally distributed. In fact, this condition is almost surely not true in the lesion segmentation mission because no raters intend to underestimate the boundary location according to potential risk consideration, especially for lesions with smooth and fuzzy boundaries, like non-pigmented lesions. In practice, some raters draw the boundary along the lesion edge strictly and produce detailed segmentations, while others prefer to include a bit of the skin region adjacent to the boundary and lead to a more compact boundary. It is obviously not fair to compare the latter (or even a compromise of both in any means that would always keep a positive distance from the truth) against the computer-based segmentation that always produces a fuzzier boundary following the pixel-wise intensity or shape variation, even though they still provide useful information about the location of the boundary. This might explain why adding a reference selection step improved the ground truth estimation result in [62].

As a result, we hypothesize that a proper **EGT** should take into account the segmentation bias pattern. In order to produce a more reasonable boundary, a prior analysis of the bias pattern would be helpful. This can later serve as *a priori* information that guides the weighting of raters and drives the **EGT** closer to the truth. This strategy has not been attempted in the related literature as far as the authors are aware.

In this Chapter, we represent the ground truth estimation as an optimization issue under a level-set framework. Two formulations are proposed based on different strategies: minimizing variation and maximizing the *a posteriori* probability. We first conduct a pattern analysis of manual segmentations and then investigate whether incorporating such pattern information will improve the ground truth estimation. For the purpose of embedding the pattern information, we will add a shape prior model term to the energy function. The performances of five different approaches will be compared by experimenting on both synthetic and real data.

Materials The 50 real skin lesion images for the purpose of testing are randomly selected from our lesion data-base. The lesion boundaries are obtained by eight dermatologists from the Dermatology department of the University of Edinburgh who directly draw the lesion boundary on the colour image displayed in Adobe Photoshop CS3 using a Wacom Clintiq 12WX Interactive pen tablet independently. We then convert the results into binary-valued images, in which the lesion is labeled as ‘1’ and the skin as ‘0’. We present the first comprehensive assessment using carefully validated segmentations of 50 lesion images. To our knowledge, ground truth estimation for lesion segmentation analysis has not been studied on such a large data set.

Notations Some notations are listed as follows:

$D_{ij}(x)$: the manual segmentation of the i^{th} image drawn by the j^{th} expert at pixel x

$T_i(x)$: the estimated ground truth of the i^{th} image at pixel x. $T_i(x) \in \{0,1\}$ and [1:

lesion, 0: normal skin]
I: the number of images
J: the number of raters
$\mathbf{P}(\Omega)$: the partition of the image Ω into N distinct regions: $\{\Omega_n\}_{n=1}^{N}$, $\cup_{n=1}^{N}\Omega_n \equiv \Omega$ and $\Omega_i \cap \Omega_j = \emptyset, \forall_{i \neq j} i, j$. Ω denotes the image domain, N is the number of regions (N = 2 for lesion and normal skin images).

3.3 Lesion Manual Segmentation Pattern Analysis

3.3.1 Impact of Different Segmentation Policies Related to Intra Rater Variation

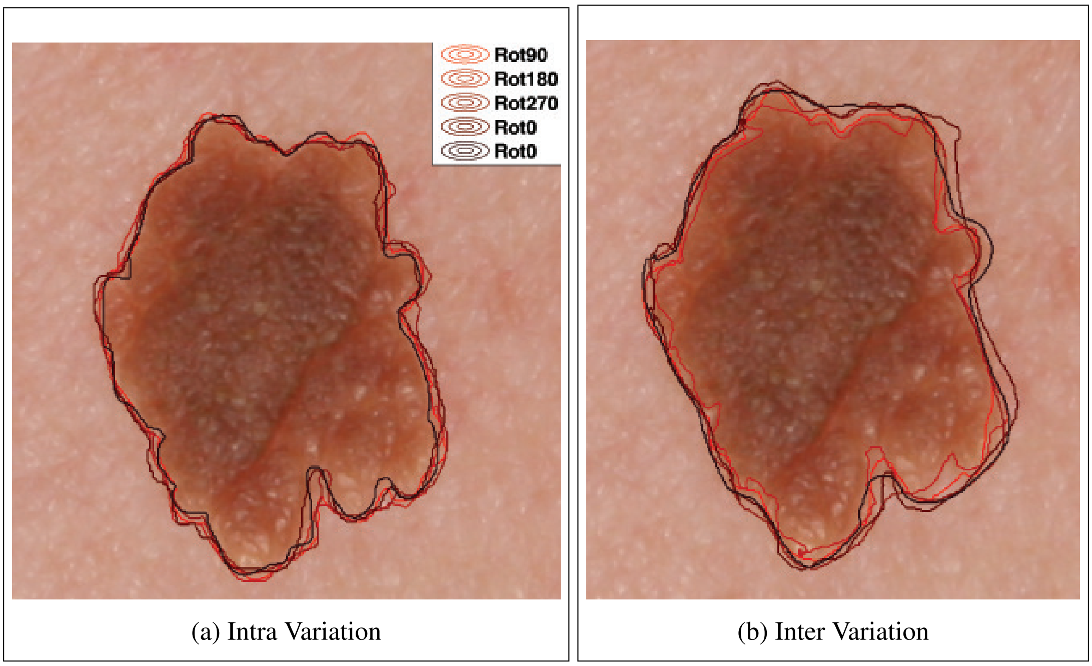
Visual inspection reveals the existence of both intra-rater and inter-rater variations on the same lesion, but the latter is more significant than the former, as shown in *Fig.* 3.1. These variations mainly take place at locations where the transition between lesion and healthy skin is smooth, *e.g.,* the blurred boundary of non-pigmented lesions and where the edge is non-convex, *e.g.,* regions where the boundary penetrates into the lesion. It would be interesting to investigate their respective impacts on the lesion labeling results.

For the intra-rater variation, one rater repeated the manual segmentation 5 times on images of the same lesion. Two trials are on the original orientation, while the other three are rotated clockwise by 90, 180, 270 degrees, respectively. As a result, we obtain 5 manual segmentations for each lesion image. The variation between each segmentation pair is measured using both *XOR* [64] and *FOM* [66].

XOR measures the spatial-region-based dissimilarity between two segmentations, *e.g.,* the real ground truth (**GT**) and the estimated ground truth (**EGT**). It has the form as:

$$XOR = \frac{Area(\mathbf{EGT} \bigoplus \mathbf{GT})}{Area(\mathbf{EGT} + \mathbf{GT})} = \frac{FN + FP}{FN + FP + 2 \times TP}. \tag{3.1}$$

It ranges from 0 (best) to 1 (worst). $\bigoplus$ denotes exclusive-OR and gives the pixels for which T_i and D_{ij} disagree; $+$ means union. The smaller the XOR, the closer the ground truth is to the manual results. The drawback of this metric is that it tends to favor larger lesions due to the size term in the denominator. FOM (Pratt's Figure Of Merit) is a distance-based measurement that is often used to compare the performance of edge detection algorithms. It stood out in comparison with five other supervised evaluation criteria for segmentation results and proved to be most effective in a comparison study



(a) Intra Variation

(b) Inter Variation

Figure 3.1: Both intra-rater and inter-rater variation exist and the latter is more significant than the former (see text for description of methodology)

conducted by Chabrier *et al.* [66]. It corresponds to an empirical contour distance between the **GT** and the **EGT** in the form of

$$FOM(CoT, CoD) = \frac{1}{max\{length(CoT), length(CoD)\}} \cdot \sum_{k=1}^{length(CoT)} \frac{1}{1+\alpha \times d^2(k)}, \tag{3.2}$$

where CoT and CoD are the boundary representations of the **GT** and the **EGT**. $d(k)$ is the Euclidean distance between the k^{th} pixel of CoT and the nearest pixel of CoD. Its weight α is set to 1 in the experiments.

The average of the measures across the 50 test images is considered as the intra-rater variation and it reflects the precision of this rater during the segmentation process. Full results are shown in *Table.* 3.1. The first row demonstrates the comparison result between the 2 non-rotated segmentations from the same person. The second row com-

Measures (×100)		XOR	FOM [66]
Intra	No rotation (2 samples)	6.33	15.66
	Rotation (4 samples)	5.80	16.67
Inter	Other dermatologist (7 samples)	8.07	12.39

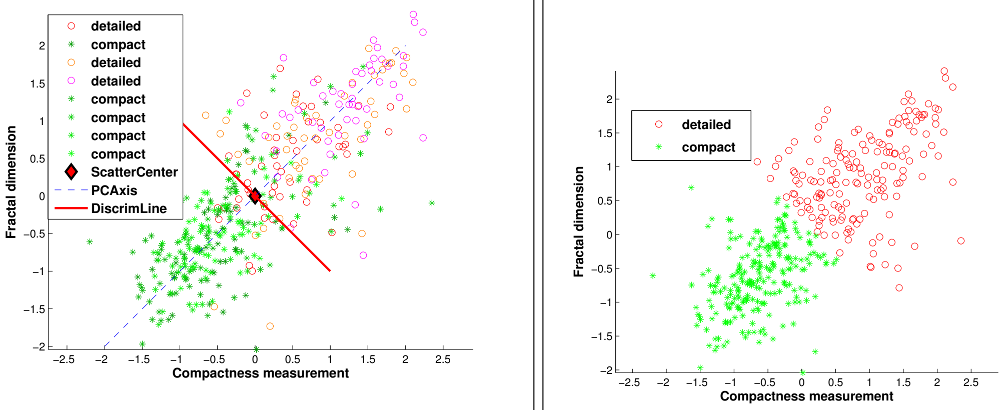
Table 3.1: Intra and Inter rater variation comparison (smaller XOR and larger FOM is better)

pares the results drawn by the same person but on 4 images rotated every 90 degrees. The third row displays the inter-rater variation, which is the average comparison results between different raters. For one thing, the intra-variation are relatively smaller compared to the inter-variation (in terms of smaller XOR and larger FOM values). On the other hand, it is possible to eliminate the intra-rater variation during the multiple results fusion process [65]. Hence, we hypothesize that *the inter-rater variation is the main factor that differentiates the segmentations from different raters and should be compensated for during the ground truth estimation procedure*.

In order to account for the inter-rater variation, the authors in [58] considered the existence of two bias patterns as underestimation and overestimation. They compensated for it through the estimated bias parameters. As discussed before, in the lesion segmentation, using these two patterns is not enough, since the rater would either trace the lesion boundary exactly, or overestimate the boundary to an extent, but no underestimation exists. Hence, the estimated ground truth would still have some overestimation since all the results make a (weighted) contribution to the ground truth. We question whether or not some other factors can help the **EGT** to converge to a more accurate lesion boundary. In the light of this, the pattern analysis of manual segmentations is necessary.

3.3.2 Two Segmentation Policy Models

Visual inspection reveals that lesion manual segmentations vary because of different raters' segmentation policies. Different opinions on the importance of finding the exact lesion boundary lead to different attitudes when people perform the manual segmentation. For some raters, locating a general lesion region is necessary for a good diagnosis. Hence, they pay less effort to the exact edge details; while other raters might pay a great deal of attention to draw a very precise pixel-by-pixel boundary. In this context, we assume that there are two patterns of manual results: segmentations that have finer details along the boundary should be comparatively more detailed and less careful segmentations that tend to have a more compact lesion region. To test this assumption, we describe all the manual results drawn by eight raters by two measurements: **Compactness measurement** ($CM = \frac{perimeter^2}{4\pi \times area}$) and **Fractal Dimension** ($FD = \frac{\log(M(s))}{\log(M)}$, M denotes the number of squares covering the image field and $M(s)$ denotes the number of squares occupied by the boundary). For each manual segmentation, a **CM** and a **FD** value are assigned. Hence, there are J manual results from different raters for each lesion i and they are denoted as $\mathbf{CM}(D_{ij})$ and $\mathbf{FD}(D_{ij})$, $i = 1,\ldots,I, j = 1,\ldots,J$. For the purpose of comparison, both **CM** and **FD** values are normalized across J raters for each lesion $(e.g.,FD_i = (FD_i - mean(FD_{i,j=1:J}))/std(FD_{i,j=1:J})))$. The scatter plot of these two values is shown in *Fig.* 3.2a. If we draw a discriminative line (*e.g.,* across the scatter plot center and perpendicular to the principal component axis direction, see dotted line in *Fig.* 3.2a), then we divide the samples into two groups. One has large CM and FD values (denoted as **detailed**) and the other has small values (denoted as **compact**). If either the **detailed** or the **compact** segmentations consistently belong to a group of raters (*e.g.,* most points on the right-hand side of the discriminative line have



(a) The discriminant line in red shows the potential separation between two pattern clusters. The segmentations with small compactness and fractal dimension values are considered as **compact** ones; while those with large compactness and fractal dimension values are considered as **detailed** segmentations. Segmentations from 8 raters are displayed in different colours. All segmentations from **detailed** style raters have reddish colour; while segmentations from **compact** style raters have greenish colour. Raters with different segmentation policies produce consistent patterns of segmentations (*i.e.,* most segmentations given by a 'detailed' style rater lays on the right hand of the discriminant line).

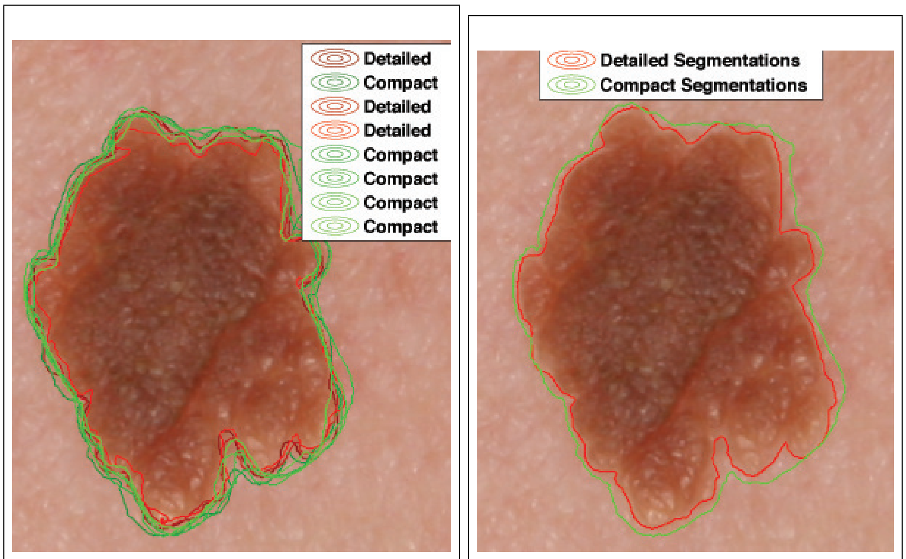
(b) All segmentations are separated into two pattern clusters using *kmeans* method. The segmentations with small compactness and fractal dimension values are considered as **compact** ones; while those with large compactness and fractal dimension values are considered as **detailed** segmentations.

Figure 3.2: The scatter plot of **FD** and **CM**.

red colour which denotes the **detailed** segmentation), then we have reasons to believe that group of raters share a common pattern of segmentation policy which differs from the other group.

In order to verify this thought, 1) we apply *kmeans* (where we force $k = 2$) to cluster all manual segmentations from 8 rater and 50 images (8×50) based on both **CM** and **FD**. Each manual segmentation result is categorized into one group (either **detailed** or **compact**). The clustering result is shown in *Fig.* 3.2b. 2) For each rater, we count the number of his/her segmentations being **compact**, as shown in the second column in *Table.* 3.2. Each rater has a corresponding cluster vector which records how compactly they draw the lesion boundary over the 50 lesions. According to the divergence between the counts for compactness, there exist two kinds of segmentation style (**detailed** and **compact**) and one can categorize raters into **detailed** and **compact** patterns based on this count (*i.e.,* if more than 25 segmentations of a rater belong to the **compact** group, this rater has a **compact** style). The categorization results are intuitively presented in *Fig.* 3.2a and *Fig.* 3.5a. Segmentations from 8 raters are displayed in different colours. All segmentations from **detailed** style raters have reddish colour; while segmentations from **compact** style raters have greenish colour. Most of the segmentations from the **detailed** raters tend to have larger CM and FD values and those from the **compact** raters tend to have smaller CM and FD values. *Fig.* 3.3b) shows ground truths estimated from **detailed** segmentations and **compact** segmentations using Majority voting based method. It can be seen that the **compact** style raters intend to include more healthy skin parts into lesion region. This result echos the rater performance parameters estimated from STAPLE (see the right hand of *Table.* 3.2), in which **compact** style raters normally have large precision but small specificity.

On the other hand, the dermatologists are reasonably consistent according to the 'counts for compact' score, which means each dermatologist obeys the same rule when doing the manual segmentation. This can be further observed from the scatter plot in *Fig.* 3.2a, in terms of the different underlying distributions of red and green samples. Therefore, it is reasonable to distinguish different raters, although the two patterns are not completely separated from each other in *Fig.* 3.2a (*e.g.,* some red / green points enter the domain of the points of the opposite colours). This is mainly because 1) there is intra-rater-variation and 2) both **CM** and **FD** values are lesion dependent, even though a normalization is performed. In order to further prove that the rater pattern categorization is feasible, we perform a statistical hypothesis test by comparing the categorization results to the hypothetical results. We suppose the null hypothesis is



(a) **Detailed** segmentations and **compact** segmentations

(b) Ground truths estimated from **detailed** segmentations and **compact** segmentations using Majority voting based method

Figure 3.3: Real segmentations

that the two patterns of segmentation (**compact** and **detailed**) for each rater are equally likely to occur, or in another words, the expected 'counts of compact' should be 25 out of 50. The question is then how likely is it to find our pattern clustering results. We consult the binomial two-tailed test $B(50, \frac{1}{2})$ to find out the probability of finding a certain number compact counts either above or below the expectation (25) in a sample of 50. The resulting P values are listed in the third column in the Table. 3.2, which indicates that raters have tendency towards one kind of pattern in segmentation and it is not random.

The above pattern analysis result can be considered as useful prior information with potential value in estimating the ground truth. In our application, we consider a good quality lesion segmentation as one which has a small average distance from the true boundary. In this context, the **detailed** segmentation outperforms the **compact** segmentation and should be considered as more important (see *Fig.* 3.3b).

	Rater Patterns Clustering based on FD and CM			**Performance(STAPLE [1])**			
Rater	counts for compact (out of 50)	Binomial Test Two-tails P Value	patterns	precision	group	specificity	group
1	18	0.0649	detailed	0.9379	small	0.9890	large
2	42	< 0.0001	compact	0.9578	large	0.9647	small
3	10	< 0.0001	detailed	0.8417	small	0.9904	large
4	3	< 0.0001	detailed	0.9095	small	0.9924	large
5	47	< 0.0001	compact	0.9466	large	0.9794	small
6	32	0.0649	compact	0.9437	large	0.9597	small
7	45	< 0.0001	compact	0.9620	large	0.9821	small
8	40	< 0.0001	compact	0.9220	small	0.9828	large

Table 3.2: Rater style clustering. *count of compact* indicates the number of compact segmentations (out of 50) produced by each rater. Based on the *count of compact* across 50 images, each rater is assigned to a pattern (*i.e.,* if a rater produces more than 25 'compact' segmentations, he/she is considered as a **compact** style rater). A binomial test shows that each rater consistently obeys his/her segmentation policy during the manual segmenting task. The performance of each rater, in terms of precision and specificity, is also estimated using **STAPLE**. Each performance parameter can be clustered into two groups ('small' and 'large') and this clustering result relates to the rater pattern clustering result (*i.e.,* **compact** style raters have large precision and small specificity because they intend to include more healthy skin into lesion region)

3.4 Ground Truth Estimation Methods

In this section, we propose two ground truth estimation algorithms: with and without taking into account a rater's performance level. Ground truth estimation is solved as an optimization issue using a level-set framework. The advantages of using a level-set framework are: 1) the force that drives the evolution of the level set function has a physical interpretation, 2) it is easy to extend to multiple category labeling problems, 3) it can also be extended to the continuous manual segmentation representation. Last but not the least, 4) the level set formalism enables us to directly incorporate prior segmentation pattern information into the estimation framework by adding a specially designed term in the energy function $E(\phi)$.

3.4.1 Variation Based Method (LSV)

The motivation behind the variation based ground truth estimation approach is to minimize the average discrepancy between estimated ground truth T_i and the manual results.

This is equivalent to minimizing the average area of the non-overlap region between T_i and D_{ij}. Hence, an energy function can be defined as:

$$\begin{aligned} E_i &= \sum_{j=1}^{J} \int_{\Omega} [T_i(x) - D_{ij}(x)]^2 dx \quad (3.3)\\ &= \sum_{j=1}^{J} \sum_{n=1}^{N} \{ \int_{\Omega_n} [T_i(x) - D_{ij}(x)]^2 dx \} \quad (3.4)\\ &= \sum_{j=1}^{J} \{ \int_{\Omega_{lesion}} [T_i(x) - D_{ij}(x)]^2 dx + \int_{\Omega_{skin}} [T_i(x) - D_{ij}(x)]^2 dx \}. \quad (3.5) \end{aligned}$$

The level-set representation of the above energy function would be:

$$\begin{aligned} E_i &= \sum_{j=1}^{J} \{ \int_{\Omega} H(\phi)[1 - D_{ij}(x)]^2 + [1 - H(\phi)][0 - D_{ij}(x)]^2 dx \}. \quad (3.6)\\ &= 2 \times \int_{\Omega} H(\phi)[\frac{J}{2} - \sum_{j=1}^{J} D_{ij}(x)] dx + \sum_{j=1}^{J} \int_{\Omega} D_{ij}^2(x) dx. \quad (3.7) \end{aligned}$$

Here, the estimated ground truth is represented by the level-set function ϕ. The boundary is the zero level-set. $H(\phi)$ denotes the heaviside step function:

$$H(\phi) \equiv H(\phi(x)) = \begin{cases} 1 & \phi(x) \geq 0, x \in \Omega_{lesion} \\ 0 & \phi(x) < 0, x \in \Omega_{skin} \end{cases} \quad (3.8)$$

$T_i(x)$ is the **EGT** label,

$$T_i(x) = \begin{cases} 1 & x \in \Omega_{lesion} \\ 0 & x \in \Omega_{skin} \end{cases} \quad (3.9)$$

Using the Euler-Lagrange equation, the minimization of $E(\phi)$ solved by a gradient descent for the embedding function ϕ is:

$$\begin{aligned} \frac{\partial \phi}{\partial t} &= -\frac{\partial E}{\partial \phi} \quad (3.10)\\ &= 2 \times \delta(\phi)[\sum_{j}^{J} D_{ij}(x) - \frac{J}{2}], \quad (3.11) \end{aligned}$$

here, $\delta(\phi) = \frac{dH(\phi)}{d\phi}$ is the Dirac delta function. It has value 1 at the lesion boundary and 0 elsewhere. From *Eq.* 3.11, the level set function ϕ tends to be stable at the position where the votes of location x as lesion and skin are tied. If more raters label x as lesion, then the zero level-set evolves towards the skin direction; otherwise, it evolves in the lesion direction. The force is determined as the distance between the overall votes of being lesion and the half of rater number. On the other hand, the energy function *Eq.* 3.7 is comprised of two terms and the second one is a constant. The energy will always be bigger than the constant unless pixels belonging to the lesion region satisfy $\sum_{j=1}^{J} D_{ij}(x) \geq \frac{J}{2}$. This is equivalent to the *Majority Vote Rule* (**MV**) theoretically. From *Eq.* 3.11, the energy function arrives at the extreme value when $\sum_{j=1}^{J} D_{ij}(x) = \frac{J}{2}$. This reveals that the *Majority Voting Rule* with voting ratio $\theta = \frac{J}{2}$ makes the estimated ground truth have the smallest average variation to the manual results. We further verify this result in Section 3.5.1.

3.4.2 Maximum *a posteriori* (MAP) Probability Based Method (LSML)

The probabilistic formulation estimates the ground truth as a process of finding an optimal partition **P**(Ω) of the image domain. It maximizes the *a posteriori* probability of a partition $p(\mathbf{P}(\Omega)|D_{i\{1,2,\ldots,J\}})$ under the condition of a set of manual results. Simply speaking, the **EGT** should be the most probable partition given all the manual results. As a result, the formulation has the form as:

$$\begin{aligned} p(\mathbf{P}|D_{i\{1,2,\ldots,J\}}) &= p(\Omega_{lesion}, \Omega_{skin}|D_{i\{1,2,\ldots,J\}}) && (3.12)\\ &= \prod_{n=1}^{2} p(\Omega_n|D_{i\{1,2,\ldots,J\}}) && (3.13)\\ &= \prod_{n=1}^{2}\prod_{x\in\Omega_n} p(T(x)|D_{i\{1,2,\ldots,J\}}(x)). && (3.14) \end{aligned}$$

$p(T(x)|D_{i\{1,2,\ldots,J\}}(x))$ is the conditional probability that pixel x belongs to region $T(x)$ (*e.g.*, $T(x) = 1$ means x belongs to the lesion) and it has the format as:

$$\begin{aligned} p(T(x)=1|D_{i\{1,2,\ldots,J\}}(x)) &= \frac{p(T(x)=1, D_{i\{1,2,\ldots,J\}}(x))}{p(D_{i\{1,2,\ldots,J\}}(x))} && (3.15)\\ &= \frac{a(x)}{a(x)+b(x)} && (3.16)\\ &= W(x). && (3.17) \end{aligned}$$

$$\begin{aligned} p(T(x)=0|D_{i\{1,2,\ldots,J\}}(x)) &= \frac{p(T(x)=0, D_{i\{1,2,\ldots,J\}}(x))}{p(D_{i\{1,2,\ldots,J\}}(x))} && (3.18)\\ &= \frac{b(x)}{a(x)+b(x)} && (3.19)\\ &= V(x). && (3.20) \end{aligned}$$

Here, a and b denotes the joint probability and are defined by assuming that the raters perform the segmentation independently:

$$\begin{aligned}
a(x) &= p(D_{i\{1,2,...,J\}}(x), T(x)=1) && (3.21)\\
&= p(D_{i\{1,2,...,J\}}(x)|T(x)=1)p(T(x)=1) && (3.22)\\
&= \prod_{j=1}^{J} p(D_{ij}(x)|T(x)=1)p(T(x)=1) && (3.23)
\end{aligned}$$

$$\begin{aligned}
b(x) &= p(D_{i\{1,2,...,J\}}(x), T(x)=0) && (3.24)\\
&= p(D_{i\{1,2,...,J\}}(x)|T(x)=0)p(T(x)=0) && (3.25)\\
&= \prod_{j=1}^{J} p(D_{ij}(x)|T(x)=0)p(T(x)=0) && (3.26)
\end{aligned}$$

$$p(D_{i\{1,2,...,J\}}(x)) = a(x) + b(x). \tag{3.27}$$

$W(x)$ and $V(x)$ are the joint conditional probability of pixel x belongs to the lesion and skin, respectively. For an individual rater, the definition of the likelihood function $p(D_{ij}(x)|T(x))$ is inspired by **STAPLE** [1] and upholds the idea that the contribution of each rater to the ground truth estimation differs based upon their performance. Hence, we have:

$$p(D_{ij}(x)|T(x)=1) = \begin{cases} p(D_{ij}(x)=1|T(x)=1) = se_j & \text{if } D_{ij}(x)=1;\\ p(D_{ij}(x)=0|T(x)=1) = 1-se_j & \text{if } D_{ij}(x)=0. \end{cases}$$

$$p(D_{ij}(x)|T(x)=0) = \begin{cases} p(D_{ij}(x)=0|T(x)=0) = sp_j & \text{if } D_{ij}(x)=0;\\ p(D_{ij}(x)=1|T(x)=0) = 1-sp_j & \text{if } D_{ij}(x)=1. \end{cases}$$

Here, $se(sensitivity) = \frac{TP}{TP+FN}$ and $sp(specificity) = \frac{TN}{TN+FP}$. Definitions of the 4 measures, TP (true positive), TN (true negative), FP (false positive), FN (false negative) are shown in *Fig.* 3.4. The prior information term $p(T(x))$ is determined solely by the labeling results at location x:

$$p(T(x)=1) = \frac{\sum_{j=1}^{J} D_{ij}(x)}{J}. \tag{3.28}$$

$$p(T(x)=0) = 1 - p(T(x)=1). \tag{3.29}$$

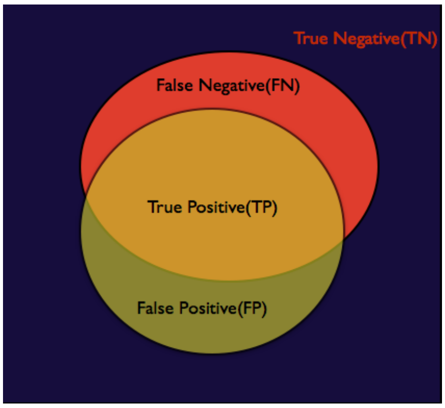


Figure 3.4: Four basic measures for binary segmentation evaluation. The red circle denotes the ground truth and the yellow denotes an estimated result

Maximizing the *a posteriori* probability in *Eq.* 3.14 is equivalent to minimize its negative logarithm as

$$E = -\sum_{n}\sum_{x\in\Omega_n} \log p(T(x)|D_{i\{1,2,...,J\}}(x)) \tag{3.30}$$

$$= -\{\sum_{x\in\Omega_{lesion}} \log p(T(x)=1|D_{i\{1,2,...,J\}}(x)) + \tag{3.31}$$

$$\sum_{x\in\Omega_{skin}} \log p(T(x)=0|D_{i\{1,2,...,J\}}(x))\} \tag{3.32}$$

According to *Eq.* 3.17 and *Eq.* 3.20, we have

$$E = -\{\sum_{x\in\Omega_{lesion}} \log(W) + \sum_{x\in\Omega_{lesion}} \log(V)\} \tag{3.33}$$

The level-set representation of the above energy function can be expressed as

$$E(\phi) = -\int_{x\in\Omega} H(\phi)\log(W) + (1-H(\phi))\log(V)dx. \tag{3.34}$$

According to the Euler Lagrange equation, maximizing of the energy functional $E(\phi)$ derives

$$\frac{\partial E(\phi)}{\partial \phi} = \delta(\phi)\left(\log\frac{W}{V}\right) = 0. \tag{3.35}$$

Solving *Eq.* 3.35 using a gradient descent for the embedding function ϕ results in a PDE which represents the contour evolution equation as

$$\frac{\partial \phi}{\partial t} = -\frac{\partial E(\phi)}{\partial \phi} = \delta(\phi)\left(\log\frac{W}{V}\right). \tag{3.36}$$

The values of W and V keep updating through iterations based on the estimated performance parameters. The physical meaning of this equation is very clear. The boundary evolves to the location where the probabilities of the pixel belonging to the lesion and the skin are identical. If the conditional probability of pixel x being lesion is larger than skin, there is a positive force proportional to $log(W/V)$ driving the boundary to move towards the skin direction and vice versa.

3.4.3 Maximum a *a posteriori* Probability Based Method Incorporating the Segmentation Pattern Information (LSMLP)

In the previous section 3.3, we categorize the manual segmentation into two patterns (**detailed** *v.s.* **compact**). Given the aim of comparing computer-based segmentations against the **EGT**, it is reasonable to generate a ground truth that has a more accurate boundary. We remark that the **detailed** segmentations suit this requirement better. Hence, we introduce a Shape Prior Model (denoted as *SPM*) that is built upon the **detailed** manual segmentations using the *Majority Vote Rule*. A shape prior based term aiming at minimizing the distance between the estimated T_i and *SPM* is formalized as:

$$\begin{aligned} E_{shape} &= \int_{\Omega}[T_i(x) - SPM(x)]^2 dx & (3.37) \\ &= \int_{\Omega} H(\phi)[T_i(x) - SPM(x)]^2 + [1 - H(\phi)][T_i(x) - SPM(x)]^2 dx & (3.38) \\ &= \int_{\Omega} H(\phi)[1 - SPM(x)]^2 + [1 - H(\phi)][0 - SPM(x)]^2 dx & (3.39) \\ &= \int_{\Omega} H(\phi)[1 - 2 \times SPM(x)]dx + \int_{\Omega} SPM^2(x)dx. & (3.40) \end{aligned}$$

We add this term to the energy function of the **LSML** in *Eq.* 3.34 and lead to a new energy function as:

$$E_{LSMLP} = E_{LSML} + \gamma \times E_{shape}. \tag{3.41}$$

Minimizing the above energy function derives the boundary evolution equation as:

$$\frac{\partial \phi}{\partial t} = -\frac{\partial E_{LSMLP}(\phi)}{\partial \phi} = \delta(\phi)\left(\log\frac{W}{V} + \gamma \times (2 \times SPM(x) - 1)\right). \tag{3.42}$$

Here, γ weights the importance of the shape prior energy (we set it to 0.4 in our experiment to give more weight on the contribution of all manual segmentations).

In such a way, we give prominence to the detailed segmentations and reduce the impact of the compact segmentations. It is worth noting that the above level-set formulations are based on two key assumptions:

1. Each rater independently performs the lesion segmentation job.

2. There is no spatial correlation between pixels.

Though the second assumption can be relaxed by incorporating a Markov random field model as stated in [58], it is out of the scope of our work.

3.5 Experiments and Results

In this section, we will compare the proposed approaches against two popular ground truth estimation methods: the *Majority Voting Rule* (**MV**) and **STAPLE**, based on both synthetic and real lesion data. For the *Majority Voting Rule* approach, we also investigate the number of agreements needed for making the decision.

3.5.1 The Best Voting Threshold

The Majority Voting Rule aims to find the ground truth with common agreement using the formulation as:

$$T_i(x) = \begin{cases} 1 & \text{if } \sum_{j=1}^{J} D_{ij}(x) \geq \theta; \\ 0 & \text{otherwise.} \end{cases} \tag{3.43}$$

Here, θ is the voting threshold that is used to determine the classification of each pixel and it is the only parameter in this simple approach. However, as pointed out by Warfield *et al.* [1], there is no guidance as to how many experts (θ) should be in agreement before making the decision. In Section 3.4.1, we showed that $\theta = J/2$ is correct choice for the estimated ground truth to have the smallest average discrepancy to the manual results. We now justify this claim with experimental evidence.

We compute the ground truth using various threshold values θ for different numbers of manual results (J). The XOR measure (mean±standard deviation) comparing the ground truth against its corresponding manual results ($\frac{\sum_{i=1}^{I}\sum_{j=1}^{J} XOR_{ij}(T_i(\theta),D_{ij})}{I \times J}$) is shown in *Table*. 3.3 (the smallest XOR measures are highlighted in red). This result shows that

XOR measure (×100)				
	Voting Threshold (k)			
Manual(J)	3	4	5	6
8	6.70±3.90	6.17±3.62	6.24±3.80	6.92±4.29
7	5.46±4.13	5.19±3.87	5.59±4.16	6.82±4.96
6	4.59±4.27	4.66±4.39	5.56±5.17	
5	3.52±3.89	4.03±4.48		

Table 3.3: Average segmentation error rates and their standard deviations

the best estimation of the ground truth is determined when using the voting method with $\theta = \frac{J+1}{2}$ when J is odd and $\theta = \frac{J}{2}$ otherwise. The latter is reasonable in practice due to the risk consideration. When a tie is reached, it is preferred to favor the more dangerous situation. The result illustrates the conclusion in Section 3.4.1. To our knowledge, we are the first group to analyse the best vote threshold for the *Majority Vote Rule* both theoretical and experimentally. Also, the XOR decreases when the number of manual results is reduced, which reflects the reduced variation among the dermatologists.

3.5.2 Comparison on Synthetic Data

In order to compare the **EGT**s derived from different approaches, it is preferred if the real ground truth (**GT**) is known. Hence, we generate synthetic data that simulates the two patterns of manual segmentations. The data is derived by using a selected computer segmentation as the ground truth that is represented as a level set function as ϕ ([67], defined based on the signed distance function from the contour). Developing the synthetic data is therefore compiled as the evolution of this ground truth. The force that drives the evolution of the level set function takes into account both systematic and random errors. The formulation of this process is as following:

$$\begin{aligned} \frac{\partial\phi}{\partial t} &= N \times F && (3.44) \\ &= N \times (Random + \nu div \frac{\nabla\phi}{|\nabla\phi|}). && (3.45) \end{aligned}$$

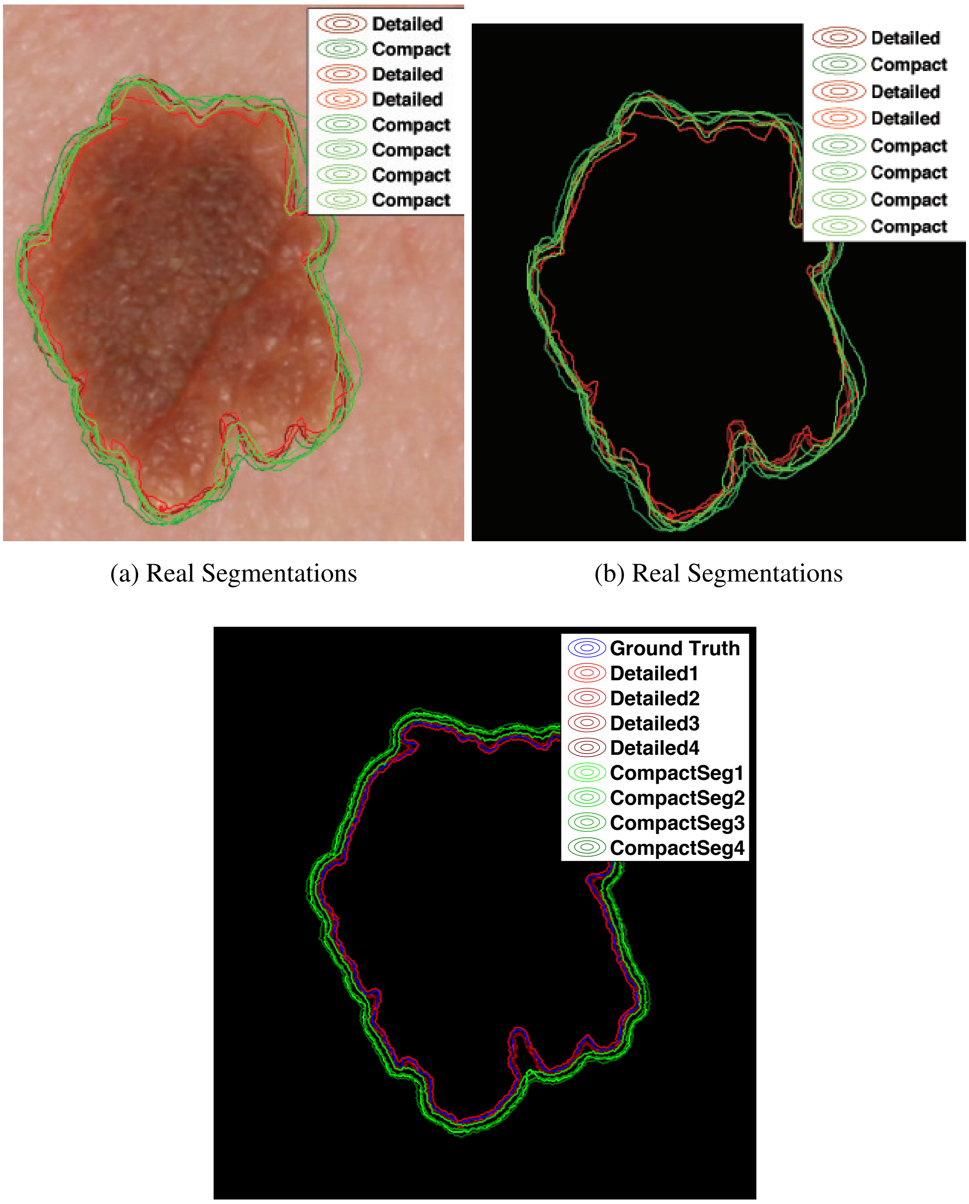
N is the normal to the curve and can be determined directly from the level set function as $N = -\frac{\nabla\phi}{|\nabla\phi|}$. F is the force and is comprised of two terms:

1. The *Random* term simulates randomness errors. A normally distributed pseudo-random value ranging from -1 to 1 is assigned to it.

2. The second term is a regularization term related to the smoothness of the evolving contour. ν denotes the weight. A larger weight is used to simulate compact segmentation; while a smaller weight is used for detailed segmentation.

Moreover, overestimation is simulated using a morphological operation: dilation. The scale of the dilation structure differs between **detailed** and **compact**, smaller for the former and larger for the latter.

The detailed segmentations are the ones with smaller ν (0.15 in the experiment) and smaller dilation scale (ranging from 2 to 5); while the compact segmentations are the ones with larger ν (0.45 in the experiments) and larger dilation scale (ranging from 4 to 10). Both are shown in *Fig.* 3.5c.



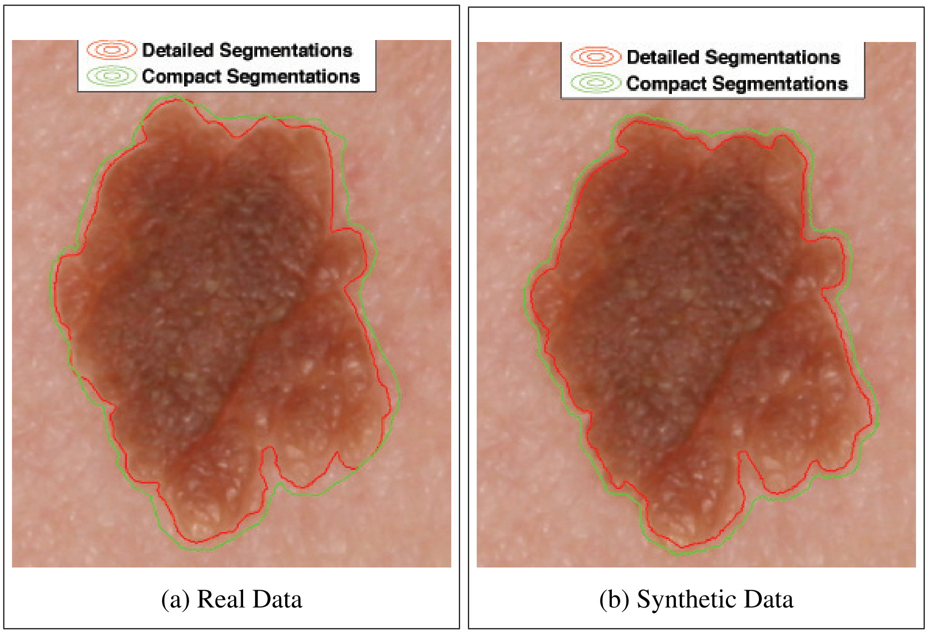
(a) Real Segmentations

(b) Real Segmentations

(c) Synthetic Segmentations

Figure 3.5: Real and synthetic segmentations

The corresponding real segmentations are shown in *Fig.* 3.5a and *Fig.* 3.5b. The detailed and compact segmentations are displayed in red and green, respectively. The ground truth estimated from each pattern of segmentations using the *Majority Vote Rule* is shown in *Fig.* 3.6. From both *Fig.* 3.5 and *Fig.* 3.6, we can see that the synthetic data have similar characteristics to the real segmentations. Hence, we use them



(a) Real Data

(b) Synthetic Data

Figure 3.6: The ground truth estimated from different patterns of the segmentations through the *Majority Vote Rule*

	Methods				
Metrics	**MV**	**LSV**	**STAPLE**	**LSML**	**LSMLP**
XOR (%)	3.8409	3.8409	3.7212	3.2733	2.1615
FOM (%)	8.9026	8.9026	10.6596	13.1484	26.7412
Sensitivity	1.0000	1.0000	1.0000	1.0000	1.0000
Specificity	0.9709	0.9709	0.9719	0.9754	0.9839

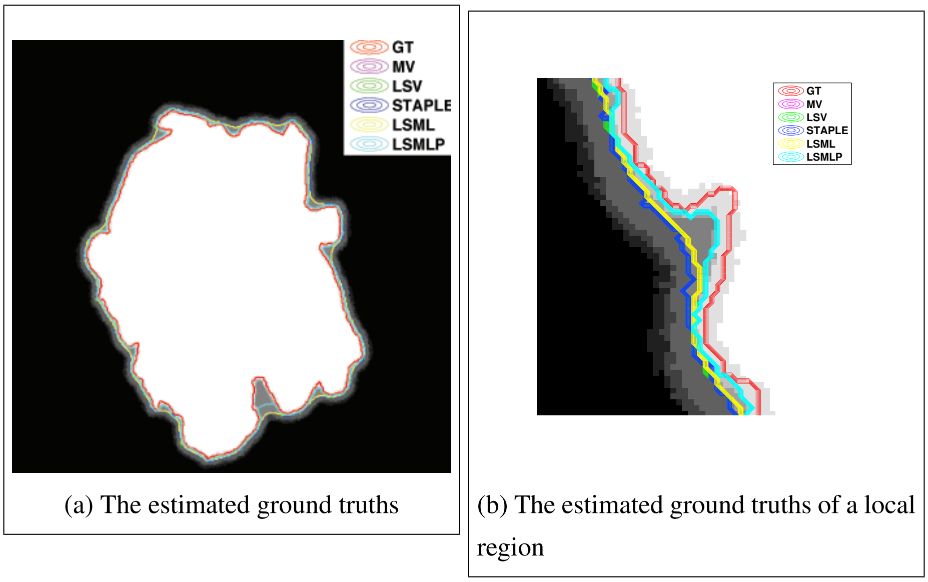
Table 3.4: The performance of different approaches according to the four metrics. The values are calculated based on the known ground truth and the ground truth estimated using different methods.

to compare different ground truth estimation algorithms. The results are shown in *Fig.* 3.7. The comparison results using *XOR, FOM, sensitivity and specificity* metrics are demonstrated in *Table.* 3.4 and *Table.* 3.5. These results show:

1. The **EGT** estimated using **LSMLP** is the closest to the real ground truth. The improvement is significant compared to the other approaches according to three metrics (*i.e.*, XOR, FOM and Specificity). The Sensitivity equals to 1 for all the methods because of the overestimation simulation. **LSMLP** outperforms the others mainly because it produces finer boundary details, especially at the locations where two groups of segmentation have big differences, such as those shown in *Fig.* 3.7b.

		Methods					
Rater Index	**Sensitivity:Specificity**	**MV**	**LSV**	**STAPLE**	**LSML**	**LSMLP**	**Real Parameters**
Rater1	Sensitivity	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	Specificity	0.9660	0.9660	0.9651	0.9616	0.9533	0.9380
Rater2	Sensitivity	0.9976	0.9976	1.0000	0.9992	0.9997	1.0000
	Specificity	1.0000	1.0000	1.0000	0.9961	0.9876	0.9719
Rater3	Sensitivity	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	Specificity	0.9737	0.9737	0.9728	0.9693	0.9609	0.9454
Rater4	Sensitivity	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	Specificity	0.9520	0.9520	0.9510	0.9476	0.9394	0.9243
Rater5	Sensitivity	0.9622	0.9622	0.9642	0.9726	0.9900	1.0000
	Specificity	1.0000	1.0000	0.9998	0.9998	0.9980	0.9858
Rater6	Sensitivity	0.8900	0.8900	0.8922	0.9002	0.9205	0.9611
	Specificity	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Rater7	Sensitivity	0.9767	0.9767	0.9768	0.9835	0.9948	1.0000
	Specificity	1.0000	1.0000	0.9991	0.9983	0.9941	0.9801
Rater8	Sensitivity	0.9613	0.9613	0.9633	0.9723	0.9942	1.0000
	Specificity	1.0000	1.0000	0.9999	1.0000	1.000	0.9861

Table 3.5: Performance level of raters estimated from different methods. The real performance level parameters for each rater are calculated from the known ground truth and the synthetic data. For each approach, the estimated performance level parameters of raters are calculated from the estimated ground truth using the corresponding approach and the synthetic data.



(a) The estimated ground truths

(b) The estimated ground truths of a local region

Figure 3.7: The ground truths estimated from the synthetic data. These grey level images are generated by aggregating individual rater binary segmentations. This provides a visual representation of rater agreement (*i.e.,* rater agreement varies with respect to grey level intensity)

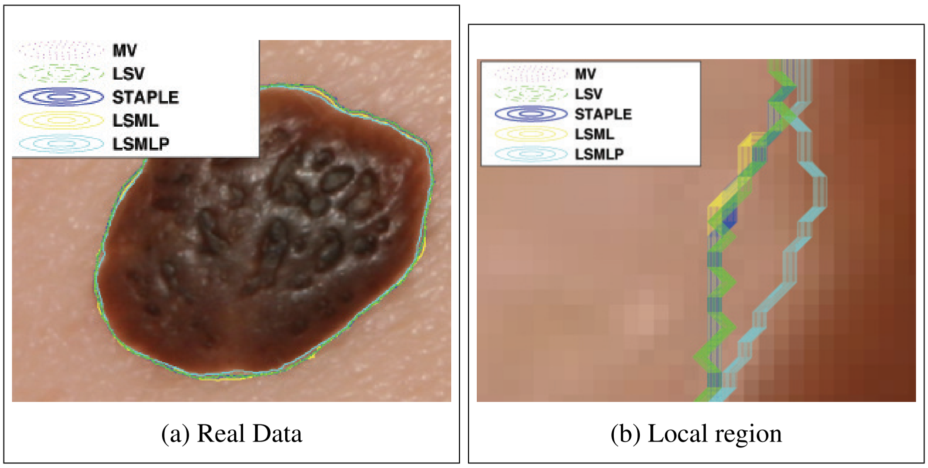
2. **LSML** produces the second best result and **STAPLE** comes the third, though there is no significant difference between them.

3. **LSV** and **MV** give the same result as predicted in our test. In *Fig.* 3.7, the boundary estimated by these two approaches overlap.

3.5.3 Comparison on Real Lesion Data

We also apply the approaches on our real lesion data, examples of which are shown in *Fig.* 3.8, *Fig.* 3.9. The same conclusion holds. The **LSMLP** outperforms the others at the locations where the boundary is non-convex and missed in the **compact** segmentations (*Fig.* 3.8b and *Fig.* 3.9b).

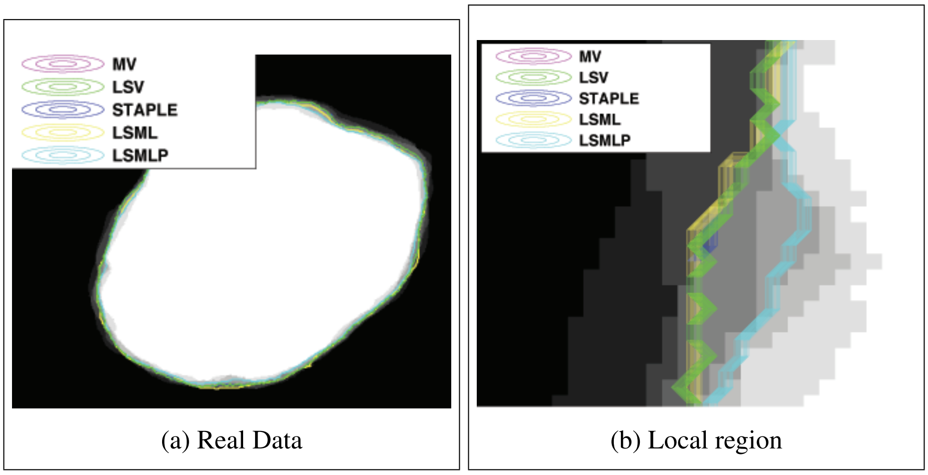
Note, for fair comparison on both synthetic and real data, the four iteration-based approaches (**STAPLE**, **LSV**, **LSML** and **LSMLP**) are initialized with the same setting: an initial circular boundary covering the lesion.



(a) Real Data

(b) Local region

Figure 3.8: Test on real data. LSV and MV give the same result as predicted in our test.



(a) Real Data

(b) Local region

Figure 3.9: Test on real data. LSV and MV give the same result as predicted in our test

3.6 Conclusion and Future work

As a means of comparing the performances of different segmentation algorithms objectively, the segmentation evaluation method needs a reference standard. But it is usually not available in real life. Therefore, estimating a **GT** from a collection of manual results becomes necessary. A good **GT** estimation algorithm should take into account both the intra and inter-rater variability, which appear in manual segmentations. We have provided evidence to argue that the latter is much more significant than the former and is the main factor that should be considered in the segmentation combination process. In order to compensate for the inter-rater variation and get reproducible results, ground truth estimation methods employ different decision fusion strategies to find a compromise between multiple manual results. In our opinion, the major difference between these approaches is whether or not they take the rater performance (in terms of weights) into account. For example, the *Majority Voting Rule* treats every rater equally; while **STAPLE** weights different raters according to the estimated performance level parameters. However, both of them ignore the characteristic of the raters' segmentations. Little research has analyzed the patterns of the manual segmentation results and we are the first group to study this subject. We found that the manual segmentations of lesion differed mainly because of the rater's segmentation policies and could be categorized into two groups: **detailed** and **compact**. Taking into account that the aim of estimating a ground truth is to compare it against computer-based algorithms, we argue it is fair to treat two patterns equally. Hence, using the categorisation result as prior information, we introduce a shape prior model that is built upon the **detailed** segmentations.

We treat ground truth estimation as an optimization problem and solve it under a level-set framework. We propose two approaches by designing the energy function based on different formulations, one is by minimizing the variation (**LSV**) and the other is by maximizing the *a posteriori* probability (**LSML**). The latter takes forward the idea in **STAPLE** [1] that takes each rater's performance level into account. The rater's pattern is incorporated by adding an energy term related to a shape prior model to **LSML** and results in a third approach called **LSMLP**. Experiments on both synthetic data and real lesion data reveal that **LSMLP** outperforms all the other methods that do not consider the prior information, followed by **LSML** and **STAPLE**.

In addition, we prove theoretically and experimentally that **LSV** and *Majority Vote Rule* (**MV**) are equivalent essentially. **MV** produces the smallest average discrepancy between the **EGT** and the manual segmentations with a voting threshold as $\theta = J/2$.

Future work will concentrate on addressing the deficiencies in the current work:

1. We generate the prior shape model by combing the **detailed** segmentations that are clustered by *k-means* using a *Majority Vote Rule* based strategy. It is worth learning the shape prior model in a more comprehensive way, *e.g.,* based on principal components analysis (PCA). Besides, other ways to incorporate the prior pattern analysis should be considered, *e.g.,*, embedding the prior information into the prior probability term $p(T(x))$.

2. Our level-set based formulations are based on an assumption that pixels have a spatial independence. We would relax this assumption by introducing a Markov

random field model as stated in [1].

3. It is of interest to extend this work into multiple phase applications, *e.g.,* multiple lesions in one image.

4. In *Eq.*3.42, γ that trades off between the LSML and shape based term was set to 0.4 arbitrarily. This is because based on the synthetic data, there is no solid ground truth and manual segmentations to estimate an optimal value for it. In any case, one could rerun the experiments using our data with varying gamma and select a value in a range of stable results. A proper way to select γ and compare different methods should be based on a test dataset with known ground truth and real manual segmentations. This data can be generated by replacing the synthetic manual segmentations with the real segmentations from dermatologists. However, the difficulty of this method lies in modifying the known ground truth (often a contour) to simulate the real-life lesion boundary which is often blurry and fuzzy.

Chapter 4

Lesion Segmentation

The chapter first presents a review of the algorithms and the information that are commonly used for skin lesion segmentation (Section 4.2). In Section 4.3, a uniform segmentation method is proposed as a test platform, which is based on a region-based probabilistic formulation of the deformable model and is implemented within the level-set framework. In order to validate the importance of depth information to lesion identification (or segmentation), diverse image cues (*i.e.,* colour, depth and texture) are incorporated into this segmentation model. Experiments and results show that the extra depth information leads to better segmentation. Section 4.4 further presents a decision tree based segmentation method that is proposed to optimize the usage of different feature combinations for segmenting different types of lesions. Experiments suggest that this strategy can further improve the segmentation performance.

4.1 Introduction

Segmentation is a mandatory step in skin lesion diagnosis, because 1) it simplifies and changes the representation of an image into something that is more meaningful and easier to analyze in the follow-up process [68], 2) the generated lesion boundary itself provides important information for accurate diagnosis and 3) the extraction of other clinical features depends on the accuracy of the boundary [69]. Any error at this stage would of course bias all the subsequent measurements and would, therefore, reduce the accuracy of the final diagnostic result [59].

Computer-based lesion segmentation plays a role as a ‘clinical eye’ that mimics and augments the dermatologist’s ability in separating the lesion from its adjacent healthy skin using various human perception-based macroscopic information and complex computer vision technologies. Compared to human segmentation, the computer-based segmentation has the advantages of being objective, consistent and efficient. Pantofaru *et al.* [70] noted that computer-based image segmentation algorithms had matured to a point that they provided segmentations which agreed to a large extent with human intuition. This might be true for common images. However, Ma *et al.* [71] argued that, due to reasons such as low contrast between structures, artifact inferences, *etc.*, the segmentations for medical applications needed more concrete background, which can be interpreted in two ways: 1) distinctive and suitable algorithms for lesion segmentation application and 2) indicative and discriminative visual information that is available to human being and useful for lesion and skin differentiation. We particularly focus on the latter as the goal of the thesis is to investigate the contribution of 3D information in lesion diagnosis.

According to the visual inspection, the lesion surface appearance is comprised of both chromatic and geometric attributes. The chromatic attribute has been commonly implied since colour information could be easily achieved by conventional 2D imaging systems. As there have been tremendous advancements in imaging techniques, higher and higher resolution imagery is becoming available [71]. Since the 1990s, the high resolution images produced by the modern imaging modality, dermoscopy (or epiluminescence microscopy) have been widely used as they offer much more lesion detail. Dermoscopy enables access to the subsurface features of lesions and therefore it can enhance the accuracy of the diagnostic and segmentation results for pigmented lesions in certain situations [72]. Most lesion segmentation methods are therefore developed for dermoscopy images and focus on pigmented melanocytic lesions (*e.g.,* melanoma and benign melanocytic nevus). Unfortunately, these methods are not suitable for the non-pigmented lesions, in which the chromatic variation between different structures is not distinguishable [52]. This type of lesions include two other important skin cancers **BCC** (Basal Cell Carcinoma, *e.g., Fig.* 4.11a) and **SCC** (Squamous Cell Carcinoma, *e.g., Fig.* 4.1a, 4.5a) for which early and correct diagnosis is also of great importance. They are included in our lesion database. For these lesions, additional diagnostic cues are desired (*e.g.,* 3D surface geometric attributes).

Compared to the chromatic attributes, the geometric attributes which reflect the topological structure of lesion surfaces have hardly been investigated, despite that they are also used as diagnostic basis by clinicians in practice. This is mainly because of the lack of corresponding 3D information which is indiscernible in the 2D images. Rigel *et al.* [72] pointed out the combination with other modalities like in-vivo range-imaging of the skin surface could provide additional features for a more reliable diagnosis. Our stereo system (see Section 2.2.1(5)) completes the story. It enables us to assess the 3D information and analyze the geometric-based features. Hence, we hypothesize this additional information could enrich the computer's vision and provide an extra attribute that helps to distinguish between lesions and their adjacent skin regions. As far as we are concerned, there has been very few contributions to the literature from both vision and dermatology sides as to whether the 3D superficial information is useful in lesion diagnosis and segmentation [47]. This makes our research meaningful.

To evaluate our hypotheses, we need to build an experimental platform - a segmentation method, which enables us to assess the impact of different lesion characteristics on the segmentation, and make a discussion on their contributions to different kinds of skin lesions (*e.g.,* pigmented, non-pigmented). We propose a segmentation approach which is suitable for our particular skin lesion application.

In the following, we will first carry out a brief literature review on segmentation algorithms as well as segmentation cues, respectively.

4.2 Segmentation Literature Review

4.2.1 Algorithms

Image segmentation refers to the process of image partition. It results in a set of segments that are adjacent and non-overlapping and collectively cover the entire image, or alternatively, a set of contours extracted from the image [68]. From a classification point of view, it can be interpreted as a process that assigns a discrete label to every pixel in an image. For a good segmentation result, all of the pixels belonging to the same region should share certain visual characteristics, such as colour, intensity, or texture; while adjacent regions should be significantly different with respect to the same characteristics. In order to achieve this goal, many algorithms using different technologies have been proposed. In the section, we will present a general review and comparison of methods belonging to the four categories: pixel-based, edge-based, region-based and deformable models based segmentation.

1. In **pixel-based** approaches, the segmentation is carried out by labeling pixels on the basis of local features like intensity. An example is the histogram thresholding segmentation, which is based on the premise that the interesting structures (*e.g.,* lesion and skin) have distinctive quantifiable features. Alternatively speak-

ing, there should exist significant peaks and valleys in the histogram, based on which the thresholds can be identified (usually found by Otsu's method). Using these thresholds, each region is comprised of the pixels whose values are within certain ranges [71]. The advantages of this method are simplicity and efficiency. It performs well in the situation where the structures have obvious intensity differences like melanocytic lesions. For instance, Xu *et al.* [73] developed the Skin Cancer Segmentation software package based on a semi-automatic method using thresholding techniques. Their experimental results on pigmented lesions showed an average error that was about the same as that obtained by experts. However, the software failed when applied to the non-pigmented lesions in the experiment conducted by Li *et al.* [22]. The failure was explained by the fact that the premise of having distinctive quantifiable intensity features in the lesion images was largely false. As it can be seen from *Fig.* 4.1a and 4.5a, the lesion region may be comprised of several kinds of tissue that result in multiple peaks and pits in the histogram; alternatively, the lesion and skin regions (like in *Fig.* 4.8a), on the other hand, may share very similar colour intensities causing the histogram to have a flat distribution. In both cases, threshold-seeking becomes difficult. Furthermore, pixel-based approaches are very sensitive to noise, such as artifacts and the uneven illumination influence. What makes things worse is that, as each pixel is processed independently, the clustering in feature space can lead to unconnected regions in image space and can also result in breaks and holes in structures. As a result, **pixel-based** approaches are not a good option for skin lesion segmentation.

2. When taking into account the local relationship between pixels, image intensity values have two basic properties: discontinuity and similarity. The **edge-based** segmentations are based on the former, in which the boundary is found where discontinuities (or abrupt changes) in features take place. Because region boundaries and image edges are closely related, the first step of **edge-based** segmentation normally involves image edge detection using various edge filtering techniques like Canny and Prewitt. Based on the filter output, pixels can be classified as edge or non-edge and pixels which are not separated by an edge are allocated to the same category [74]. The most popular and advanced **edge-based** segmentation approach is the 'Snake' or active contour model [75]. It also belongs to the **deformable model** based segmentation category that we will

discuss later. The 'snake' is in fact a spline that moves within images to find object boundaries. It is pulled towards the object boundary by both internal (*e.g.,* spline bending) and external forces. The external forces are image related and are often generated based on the image gradient. Recently, a new type of Snake that is referred to as the Gradient Vector Flow (GVF) Snake has become attractive in the field. The movement of the GVF Snake is determined by a field of forces that is calculated as a spatial diffusion of the gradient of an edge map derived from the image. Tang [76] presented a segmentation algorithm using a multi-directional GVF Snake and applied it on the pigmented skin lesion. They concluded that the performance of the approach was close to human segmentation. Zhou *et al.* [77] combined local GVFs with a mean shift strategy to derive a dynamic energy force for snakes. The experiments showed that their method was capable of accurately determining skin lesion borders in dermoscopy images. However, it is not clear which kind of lesion information (*e.g.,* greyscale intensity or colour) was used in the segmentation scheme. The foundation of the **edge-based** segmentation is to utilize image gradient related information to identify object boundaries. This kind of method works well when there is a sharp variation in intensity at the region boundaries, but it fails when the boundary between structures is blurred. The latter is a common situation for non-pigmented lesions. On the other hand, as being highly localized image information, the image edge has been found to be very sensitive to image noise [78]. Moreover, it is not convenient to incorporate multiple image cues in the **edge-based** segmentation. As a result, neither of the **pixel-based** nor **edge-based** segmentations are competent for all skin lesion segmentations.

3. The **region-based** approaches come from the observation that quantifiable predefined features inside a structure tend to have visual similarity and strong statistical correlation. The main advantage of the **region-based** segmentation in contrast to the **pixel-based** and **edge-based** segmentation is that it tends to be less sensitive to noise. This explains why it is the most popular method for medical image segmentation [71]. For particular applications like skin lesions whose edge information is ambiguous, partitioning using **regional-based** scheme seems to be more appropriate. In general, the procedure of **region-based** segmentations involves: 1) choose a proper set of features which can identify the same-content regions and simultaneously differentiate different-content regions and 2) apply a

segmentation model to the chosen features to achieve a segmentation map [79]. The most commonly used segmentation model is based on clustering techniques, which perform clustering in a feature space. This kind of method normally operates iteratively by grouping together pixels which are neighbours and have similar values and splitting groups of pixels which are dissimilar in value [74]. Commonly used algorithms are region growing, Watersheds [80] and k-means, *etc.* In [81], Iyatomi *et al.* proposed a dermatologist-like lesion region extraction algorithm based on the region-growing approach and brought the extraction results closer to those determined by dermatologists for both XLM (oil immersion and cross polarization mode of epiluminescence microscopy (ELM)) images and TLM (side transillumination mode of ELM) images. Yuan *et al.* [82] proposed a novel multi-modal skin lesion segmentation method based on region fusion and narrow band energy graph partitioning. Comparisons showed that their method outperformed the state of the art methods. Their approach only used intensity features and an extension to incorporate colour and texture features was considered as future work.

4. More recently, **deformable models** (or active contours) have been intensively investigated and applied to segment medical images because they are very flexible and can be used for complex segmentations and produce promising results [83, 84, 71]. The basic idea is to allow the contour to deform so as to optimize a given energy function and the structure boundary is therefore the final status of the initial contour. In medical application, the **deformable models** based segmentations have been particularly focused on region-based flows, because of many advantages when compared to edge-based methods including robustness against initial curve placement and insensitivity to image noise [71]. While the edge-based active contour (*e.g.,* Snakes) is evolved by fitting to local edge information, the region-based deformable model attempts to fit models to intensity, colour, texture or other sophisticated shape and appearance features within each of the separated regions and finding an energy optimum where the model best fits the image. Some of the most well-known and widely used region-based active contour models assume the various image regions to be of constant intensity. For example, the Chan-Vese algorithm used the classical Mumford-Shah functional and found a uniform smooth approximation of each region [85]. However, as one can see, the appearances of structures in medical images, *e.g.,* skin lesions,

are usually highly textured and have considerable variances within the same region (as in *Fig.* 4.1a). Therefore, although most feature extraction methods are designed to extract a uniform response for all pixels in one class, the features often vary due to a non-stationary distribution of pixels in the same region or the existence of noise in the image [79]. Hence, unless objects have distinct features (*e.g.,* colours) and are well separated, otherwise, this type of method can encounter problems when there are many complex objects with less distinct features or there exists gradual variation in colour, illumination, shading and textures [78]. As a result, more advanced techniques should be considered. One of the most successful developments is to incorporate statistical region-based models. This method attempts to model regions by known distributions, intensity histograms, texture maps, or structure tensors. In practice, the method also enables the incorporation of professional knowledge such as anatomical structures and the spatial relationship [71]. As a result, we choose to perform the skin lesion segmentation using the statistical region-based deformable model.

The key to the **deformable model** is the energy function, based on which the contour evolves to an optimum. According to the way that different functions are built, deformable models can be further classified as parametric and geometric models [71]. Currently, the latter tends to replace the former because it is able to resolve several drawbacks encountered in the parametric models, such as failing when undergoing topological changes. More details can be found in [86].

For building the geometric models, we have chosen to look at the level-set framework which has become widely used in the vision community [70]. In the subject of segmentation, except for dealing with the topological changes problem, it also has these advantages: 1) being generally effective, 2) yields an informative representation of regions and their boundaries on the pixel grid without the need of complex data structures, 3) simplifies the optimization for deformable models as standard numerical methods can be employed and 4) increases the flexibility of the deformable model and allows the use of various kinds of features, shape knowledge, *etc* [87]. In the level-set framework, contours are implicitly represented as the (zero) level-set of some embedding function (*e.g.,* ϕ). For the statistical region-based segmentation, one can define a cost function (or an energy function) that reflects regional forces based on parametric models of features (often a Gaussian because of mathematical tractability). The corresponding Partial

Differential Equation (PDE) of the function is then derived using the Euler equations. Based on the PDE, which can be solved using a gradient descent method for the embedding function ϕ, the contours evolve in the direction of a negative energy gradient [86]. The parametric model of PDEs and the segmentation are computed jointly.

4.2.2 Segmentation Information

Since a computer-based lesion diagnostic system plays the role of a 'clinical eye' that mimics the clinician, a good 'clinical eye' is fundamental. However, as current segmentation approaches are refined and new techniques are developed, one stumbling block that remains is the lack of comprehensive vision information. *Our concern is how can computers make good and comparative decisions to those given by human being if they are not provided with the same source of information?* We believe that the input to the computer-based segmentation systems should be human perception-based information and have the characteristics of correctness and comprehensiveness.

Grey Value

So far, grey value-based descriptors have been widely used. Some of them are certain component of a colour space. Others are obtained by converting a colour image to greyscale. For example, in [88], the segmentation is based on a greyscale image that is calculated as the lightness component of the HSL colour space. In [35], the author introduced two segmentation methods that were operated on a single colour channel (*e.g., R*) or a greyscale channel derived from the RGB colour space which was preprocessed to enhance the colour contrast. The lack of discriminative cues obviously limits the power of segmentation algorithms of producing high quality results. To increase the discriminative power, other descriptors like colour, depth and texture should be taken into account [89].

Colour

It is well known in the field that colour information is very important for the visual and the computer-aided diagnosis of melanoma [36] as well as for other lesion types. Colour images contain far more information than grey-scale images. They enable a more accurate and detailed assessment of the appearance of lesions and therefore, they should lead to a higher quality segmentation [90]. The question emerges as how to represent colour information in terms of features. Generally, colour features can be extracted as different channels from various colour spaces, such as *RGB*, *HSV* and

CIE_Lab. *RGB* is an obvious choice as it needs no transformation and is directly acquired by the imaging system. Particularly, skin lesions are often more prominent in channel *B* [69]. But one problem of *RGB* is that each component often has high correlation with the others. On the other hand, their absolute values are environmentally dependent. Both *CIE_Lab* and *HSV* are nonlinear transformations of the RGB colour space. Their common advantages are 1) they yield perceptually uniform spacing of colours, and show consistency to human vision systems, 2) They are capable of handing brightness and chromaticity information separately. By discarding the brightness component it is possible to make the analysis independent from the intensity variations of the environmental illumination and achieve the actual colour information. For example, in *HSV* colour space, *V* describes the intensity of brightness. It is independent from the chromatic ones; while *H (hue) and S (saturation)* colour models are scale-invariant and shift-invariant with respect to the light intensity. In addition, these two components are closely related to the way in which the human visual system (HVS) perceives colour [36, 91]. However, no dominating advantage of one colour space or one channel has yet been found [90]. The choice of colour space and colour channel should be application dependent. Garnavi *et al.* [31] proposed an automatic border detection method which integrated the optimal colour channels selection. The optional colour channels were 25 colour features extracted from six colour spaces. Their segmentation performances were evaluated by comparing to the dermatologist-drawn borders based on various metrics (*e.g.,* accuracy, precision, sensitivity, specificity, and border error). The optimal colour channels were chosen as the input to a hybrid thresholding based segmentation approach. The final border was the combination of the segmentation results obtained using individual colour channels based OR operation. The authors claimed that the method was highly competitive with three state-of-the-art border detection methods and potentially faster, since it mainly involved scalar processing as opposed to vector processing performed in the other methods. However, as the colour feature selection is simply based on the ranking results without taking into account the correlation between individual colour channels, it is doubtful that the combination of the segmentation results using those chosen colour features based on OR operation would produce an overall optimal segmentation result.

Depth

Current lesion segmentation algorithms are mostly designed for pigmented lesions, *e.g.,* melanoma and melanocytic nevus. For these lesions, the chromatic attributes are often adequate for distinguishing different regions. However, if we consider a broader lesion field like non-pigmented lesions, the chromatic attributes are obviously far from sufficient as most of these lesions do not have distinct colour features. Therefore, features from other cues are worth exploring.

Visual inspection reveals that most lesions have observable superficial geometric variation, *e.g.,* elevation from the adjacent skin. As it can be seen from *Fig.* 4.7c, there is significant shape variation in the 3D data, but the corresponding left-hand region does not have correlated colour variation, as shown in *Fig.* 4.7a and 4.7b. From the regional distributions in *Fig.* 4.7d, one can see that the two regions can hardly be distinguished from each other along the x-axis (representing colour property), but they are more separable in the y-axis (representing depth property). Unfortunately, because of the lack of 3D imaging system, segmentations using lesion surface shape information has been little investigated to the best of our knowledge.

Texture

Texture is another important property for skin lesions since many of them are highly textured (*e.g.,* Seborrheic Keratosis) and the underlying texture property distributions are different for different regions (*e.g.,* lesion and healthy skin). The problem is that texture attribute extraction is a fairly difficult topic. There is even no clear definition of what texture is. Therefore, no principled answers exist on how to texture-label pixels, but there are some *ad-hoc* suggestions [92]. The most widely used texture features are filter based, *e.g.,* the ones extracted using the Gabor filters. However, these filter-based features have the common drawback of leading to a high dimensional feature space. In practice, handling large dimensional data is difficult. Also, there is usually a significant amount of redundancy among these filtering responses [93]. Even though the dimensionality problem can be resolved using some dimension reduction operation like Principal Component Analysis (PCA), these operations themselves can introduce additional complexity *e.g.,* parameters need to be tuned. As in lesion images, the local repetition of the lesion structure provides the basis for the appearance of a texture pattern in the neighborhood region, the co-occurrence matrices based feature is also considered in representing lesion texture. In [94], Dhawan *et al.* described a multichannel segmentation algorithm which used both grey-level intensity and co-occurrence matrices based features for region extraction. They concluded the incorporation of grey-level intensity and texture feature produced better result than that obtained using the grey-level intensity feature alone. However, as a separate co-occurrence matrix has to be calculated for each pixel based on pixels around the original pixel, the calculation of textural feature is very computational expensive. An alternation could be the nonlinear structure tensor based descriptor which only involves the calculation of the first partial derivatives at each pixel and produces good properties for texture discrimination. This textural descriptor has become popular in texture representation recently [86].

In the next section, we will give a detailed description of the implementation of our segmentation approach, covering the general formulation, lesion descriptors and some variants of statistical region models. In order to distinguish this first algorithm from the latter development based on a hierarchical strategy (in Section 4.4), we call it ‘Uniform segmentation’.

4.3 Uniform Segmentation

Through the literature review and concrete visual analysis of skin lesions, we have chosen to utilize the **deformable model** based segmentation algorithm and implement it within a level set framework. The corresponding energy function is built upon the statistical region-based models.

4.3.1 Method

From a pattern recognition point of view, segmentation is also a classification problem. Hence, it is convenient to borrow the concepts from Bayesian inference theory which estimates the conditional probability of a hypothesis being true based on some forms of evidence. The energy function of the **deformable model** is designed in a probabilistic formulation using Bayesian inference theory. Mathematically, it is equivalent to maximize the *a posteriori* (MAP) probability. The segmentation is found as the most possible partition, T, of the image domain Ω that maximizes the conditional probability of lesion information I. The general formulation is:

$$T^* = \arg\max_{T\in\Omega} p(T|I). \tag{4.1}$$

According to the Bayesian Chain Rule, the *a posteriori* probability can be further expressed as

$$p(T|I) = \frac{p(I|T)\cdot p(T)}{p(I)}, \tag{4.2}$$

where, $p(I|T)$ is the likelihood function, referring to the conditional probability of the observation of I. The prior probability of the segmentation that is inferred before any evidence becomes available is represented by $p(T)$. The marginal probability of image observation is denoted as $p(I)$. Because it does not vary with respect to any solutions of T, therefore, it is considered as a constant and can be ignored. This leads to:

$$p(T|I) \propto p(I|T) \cdot p(T). \tag{4.3}$$

As maximizing the *a posteriori* probability is equivalent to minimizing its negative logarithm, our energy function can be further expressed as:

$$E = -\log\{p(I|T) \cdot p(T)\} = -\log p(I|T) - \log p(T) = E_1 + E_2. \tag{4.4}$$

In the following, we will specify the individual terms in *Eq.* 4.4, respectively.

1. **Image Based Term**

   E_1 is an image based term that corresponds to the likelihood function:

   $$E_1 \;=\; -\log p(I|T). \tag{4.5}$$

   One can assume that: 1) the image partition T is composed of two non-overlapping regions $\{\Omega_1, \Omega_2\}$ (where $\Omega_1 \cup \Omega_2 \equiv \Omega$ and $\Omega_1 \cap \Omega_2 = \emptyset$. Note, we only consider the binary skin lesion segmentation), 2) no correlation between labelings (*i.e.,* pixel-wise independent assumption) [1] and 3) values at different locations of the same region can be modeled as independently and identically distributed realizations of the same random process. Based on these assumptions, the likelihood function can be extended as following [86]:

   $$\begin{aligned} p(I|T) &= \prod_{i=1}^{2} p_i(I|\Omega_i) & (4.6)\\ &= \prod_{i=1}^{2} \prod_{x\in\Omega_i} p_i(I(x)|\Omega_i). & (4.7)\end{aligned}$$

   Hence,

   $$\begin{aligned} E_1 &= -\log \prod_{x\in\Omega_1} p_1(I(x)|\Omega_1) - \log \prod_{x\in\Omega_2} p_2(I(x)|\Omega_2) & (4.8)\\ &= -\sum_{x\in\Omega_1} \log p_1(I(x)|\Omega_1) - \sum_{x\in\Omega_2} \log p_2(I(x)|\Omega_2). & (4.9)\end{aligned}$$

   The level-set formulation of the above equation is:

   $$E_1(\phi) = -\int_{x\in\Omega} [H(\phi(x)) \log p_1(I(x)|\Omega_1) + (1 - H(\phi(x))) \log p_2(I(x)|\Omega_2)] dx. \tag{4.10}$$

[1] Further discussion on this assumption is given in the next section: **Prior Information Based Term.**

In *Eq.* 4.10, the two terms model the areas inside and outside the lesion boundary, respectively. p_1/p_2 are their corresponding probability density functions (*pdf*s). $H(\phi)$ denotes the heaviside step function,

$$H(\phi) = \begin{cases} 1 & \phi(x) \geq 0, x \in \Omega_1, T(x) = 1 \\ 0 & \phi(x) < 0, x \in \Omega_2, T(x) = 0. \end{cases} \tag{4.11}$$

Considering the associated Euler-Lagrange equation for ϕ, the minimization of the energy functional by a gradient descent of the embedding function ϕ is:

$$\frac{\partial \phi(x)}{\partial t} = -\frac{\partial E_1(\phi)}{\partial \phi} \tag{4.12}$$

$$= \delta(\phi(x)) \left(\log \frac{p_1(I(x)|\Omega_1)}{p_2(I(x)|\Omega_2)} \right). \tag{4.13}$$

$\delta(\phi) = \frac{dH(\phi)}{d\phi}$ has value 1 at the lesion boundary and 0 elsewhere. According to *Eq.* 4.13, the level set function ϕ tends to be stable at the position where the likelihood probabilities corresponding to the lesion and skin models are equal. If the likelihood probability at x is larger under the lesion-based model, then the zero level-set evolves towards the skin direction so that x is included in the lesion region; otherwise, $\phi(x) = 0$ evolves in the lesion direction and x is included in the skin region. From the pattern recognition point of view, this procedure is similar to a supervised probabilistic classification. At each iteration, the training data are related to the current position of the contour. The two classes inside and outside the contour are modeled as p_1 and p_2. The test data are the ones on the lesion boundary. Their belongings are determined by the discriminative function $\log \frac{p_1(I(x)|\Omega_1)}{p_2(I(x)|\Omega_2}$. Based on this categorization result, the components of different classes in the training data are adjusted at the end of each iteration. This is reflected as the involvement of the boundary in segmentation. The iteration will not stop till the contour evolves to the discriminative boundary.

There are two concerns about *Eq.* 4.13: 1) how to choose a statistical model (p_i) to fit the density distribution of image information I and 2) how to represent the image information in terms of features or properties $f(x)$. They will be addressed in details in Section 4.3.2 and Section 4.3.3, respectively.

2. **Prior Information Based Term**

   $E_2 = -\log p(T)$ denotes a prior information related term. Ma *et al.* [71] proposed that information used for lesion image segmentation did not only come from lesion appearances but also from some prior knowledge. Hence, this term is important for designing an effective algorithm. Taking into account prior information will allow us to cope with the missing low-level information and appear to be especially helpful in the applications where the data are influenced by noise or partial volume effects. Two kinds of priors are popular: the generic priors (also called geometric priors) and the object specific priors [86]. The former are commonly represented as a regularization constraint that favors a short length of the contour. The latter often incorporate a shape model that is statistically learnt from a set of samples. However, in our work, we will use this term to model the label distribution so as to relax the second assumption made in *Eq.* 4.7. In *Eq.* 4.7, it is assumed that the likelihood probability is independent of the neighborhood at each pixel location, but in practice, this assumption is not valid because it is often the case that the segmentation (denoted as T) has underlying spatial correlations [1].

In order to model the spatial correlations, we assume that the random field T fulfills a Markov condition with respect to the local region. According to the Hammersley-Clifford theory, the Markov Random Field (MRF) T can be modeled in the form of the Gibbs distribution as the following [79]:

$$p(T) \quad = \quad \frac{1}{Z}\exp\{-\frac{1}{\tau}U(T)\}, \tag{4.14}$$

where, τ is a positive constant and Z is defined in the form:

$$Z = \sum_{T}\exp(-\frac{U(T)}{\tau}). \tag{4.15}$$

$U(T)$ is an energy function that incorporates the neighborhood relationship (depends solely on the pairwise interactions) and can be defined as

$$U(T) = -\sum_{x\in\Omega}\sum_{y\in N_x} V(T(x),T(y)), \tag{4.16}$$

where, y is within the neighboring region of x (N_x). When embedding the local homogeneity relationship into V, we have

$$V(T(x),T(y)) = \gamma_{xy}\times(T(x)T(y)+(1-T(x))(1-T(y))). \tag{4.17}$$

Here, γ_{xy} weights the contribution of the neighborhood pixel y to the position x. Normally, larger values are assigned to the closer neighborhood pixels, with smaller values assigned to the distant pixels. As a result,

$$\begin{aligned} E_2 &= -\log p(T) && (4.18)\\ &= -\log\{\frac{1}{Z}\exp\{-\frac{1}{\tau}U(T)\}\} && (4.19)\\ &= \frac{U(T)}{\tau} + \log Z && (4.20)\\ &= -\frac{1}{\tau}\sum_{x\in\Omega}\sum_{y\in N_x} V(T(x),T(y)) + \log Z. && (4.21) \end{aligned}$$

Because Z is defined over all possible configurations on T and is impractical to evaluate, it can be considered as a constant like τ [1, 79]. Hence, we have:

$$\begin{aligned} E_2 &= -\frac{1}{\tau}\sum_{x\in\Omega}\sum_{y\in N_x} V(T(x),T(y)) && (4.22)\\ &= -\frac{1}{\tau}[\sum_{x\in\Omega_1}\sum_{y\in N_x} V(T(x)=1,T(y)) + && (4.23)\\ &\quad \sum_{x\in\Omega_2}\sum_{y\in N_x} V(T(x)=0,T(y))]. && (4.24) \end{aligned}$$

Its level-set representation can be expressed as

$$\begin{aligned} E_2 &= -\frac{1}{\tau}\int_{x\in\Omega}\{H(\phi(x))\sum_{y\in N_x} V(T(x)=1,T(y)) + && (4.25)\\ &\quad (1-H(\phi(x)))\sum_{y\in N_x} V(T(x)=0,T(y))\}dx && (4.26)\\ &= -\frac{1}{\tau}\int_{x\in\Omega}[H(\phi(x))\sum_{y\in N_x}\gamma_{xy}\times T(y) + && (4.27)\\ &\quad (1-H(\phi(x)))\sum_{y\in N_x}\gamma_{xy}\times(1-T(y))]dx && (4.28)\\ &= -\frac{1}{\tau}\int_{x\in\Omega}[H(\phi(x))\sum_{y\in N_x}\gamma_{xy}\times(2\times T(y)-1) + && (4.29)\\ &\quad \sum_{y\in N_x}\gamma_{xy}\times(1-T(y))]dx. && (4.30) \end{aligned}$$

The evolution equation of the level set function $\phi(x)$ is derived as:

$$\frac{\partial\phi(x)}{\partial t} = -\frac{\partial E_2(\phi)}{\partial\phi} = \frac{1}{\tau}\times\delta(\phi(x))\left(\sum_{y\in N_x}\gamma_{xy}\times(2\times T(y)-1)\right). \quad (4.31)$$

The above equation only takes into account the pairwise spatial interaction. Taking the simplest situation where weights γ_{xy} are uniform for instance, the iteration terminates at the segmentation boundary where the labeling for lesion and skin in the neighborhood region of x are identical. If there are more lesion labelings than skin, it indicates that x has a larger chance of being a lesion pixel.

Therefore, a force proportional to $2 \times T(y) - 1$ will drive the zero level set to move towards the skin region and vice verse.

In our implementation, we simplify the model by defining the neighborhood N_x of each pixel position x as a circular area with radii of 5. All the weights γ_{xy} to individual neighborhood pixels are set to 1.

The complete energy function is equivalent to

$$\frac{\partial \phi(x)}{\partial t} = \delta(\phi(x)) \left(\log \frac{p_1(I(x)|\Omega_1)}{p_2(I(x)|\Omega_2)} + \frac{1}{\tau} \times \sum_{y \in N_x} (2 \times T(y) - 1) \right). \tag{4.32}$$

For the convenience of further analysis, we modify the *Eq.* 4.32 by replacing the weight parameter $\frac{1}{\tau}$ with β, which now weights the image based term. Hence, the final level-set evolution equation is:

$$\frac{\partial \phi(x)}{\partial t} = \delta(\phi(x)) \left(\beta \times \log \frac{p_1(I(x)|\Omega_1)}{p_2(I(x)|\Omega_2)} + \sum_{y \in N_x} (2 \times T(y) - 1) \right). \tag{4.33}$$

In *Eq.* 4.33, the only parameter that needs to be determined is weight β. It adjusts the contributions of the image information based component and the regional labeling component to the whole system. β is normally set as a constant, but the authors in [79] argued that inappropriate setting of this constant can result in three consequences: 1) inaccurate use of regional image information using small β, 2) ignorance of the prior information based term like spatial relationships with large β and 3) the result converges to a locally but not globally optimal solution when using a balanced β. They resolved the issue by introducing a variable weighting parameter $\beta(t) = c_1 \times 0.9^t + c_2$. This implementation scheme may enable the system to converge to a global optimal (*i.e.,* an accurate estimation of regional models) at the beginning and then refine the result by taking into account the spatial relationship information. Another advantage of this strategy is it makes the segmentation less sensitive to the initial contour. In the experiment, we set $c_1 = 80$ and $c_2 = 1$ empirically ($\beta(t)$ gradually converges to $\beta = 1$).

4.3.2 Image Properties

The essential factor for the success of lesion segmentation is how well the features can characterize lesion and distinguish it from the adjacent skin. Hence, the question arises as which kind of image cues can be used to represent the lesion image I and how to describe them in terms of features $\{f_i\}$.

In this section, we will group the human-perception based information into 3 categories: colour, depth and texture and discuss them respectively.

Colour

In our work, instead of using features from a particular colour space, we consider all three colour spaces: *RGB*, *HSV* and *CIE_Lab*, though only colour descriptors that are invariant to changes of lighting conditions are considered. This is because of the concern that medical images are often severely affected by lighting conditions [89]. In this context, L and V in the *CIE_Lab* and *HSV* colour spaces do not qualify because they directly relate to the illumination parameter. For the properties in the *RGB* colour space, we perform a normalization operation to remove the influence of the illumination variation and obtain three chromatic-based substitutions as r, g, b (*e.g.,* $r = \frac{R}{R+G+B}$). On the other hand, it is worth noting that H in the *HSV* colour space cannot be directly used as a feature referring to a pure colour. Because a hue is an element of the colour wheel, which starts at red primary at 0° and wraps back to red at 360°. Thereby, similar colours belonging to the red category can be assigned to two extremely different values along the intensity interval (*e.g.,* one as 0°, whereas the other as 360°). This is particularly tricky for skin lesion colour representation as the majority lesion pixels have the hue of red. To solve this problem, we modify the hue channel using a shifting method, the details of which can be found in Appendix A. We denoted the modified hue as $\widehat{H}$. Hence, the full list of colour feature set is

1. the chromacity component r from the normalized *RGB*.
2. the chromacity component g from the normalized *RGB*.
3. the chromacity component b from the normalized *RGB*.
4. the modified Hue ($\widehat{H}$) of *HSV*
5. the Saturation (S) of *HSV*
6. the a^* of *CIE_Lab*
7. the b of *CIE_Lab*.

In order to reduce the feature dimension and avoid the redundancy between colour-based properties, we propose a feature selection procedure which is detailed in Section 4.4.3. The final choice of colour-based properties are

1. the modified Hue ($\widehat{H}$) of HSV

2. the Saturation (S) of HSV

3. the a^* of CIE_Lab

4. the chromacity component b from the normalized RGB.

For each colour feature image, image smoothing based on adaptive anisotropic diffusion (detailed in the texture section) is applied to reduce the influence of artifacts like hair and intrinsic cutaneous features (*e.g.,* blood vessels, skin lines). In this context, each image position x is associated with a colour-valued feature vector, as $I_{colour}(x) = (f_{hue}(x), f_{saturation}(x), f_{a*}(x), f_{blue}(x))^T$. As shown in *Fig.* 4.1, 4.2, the lesion areas are enhanced compared with the conventional RGB representation for both pigmented and non-pigmented cases. In addition, the colour space transformations help to remove the influence of the imaging noise arising from specular reflection. For example, the high reflection spots in *Fig.* 4.1a, 4.1b, 4.1c and 4.1d are removed in *Fig.* 4.1e, 4.1f, 4.1g and 4.1h. We refer to these four colour-based properties as C.

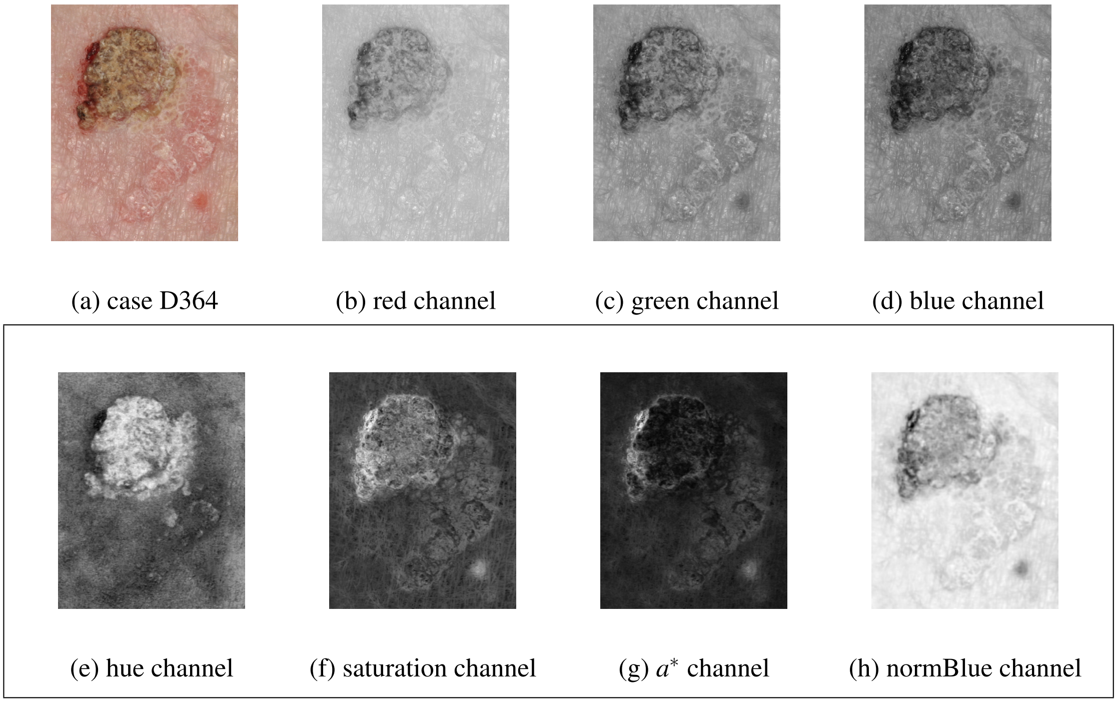
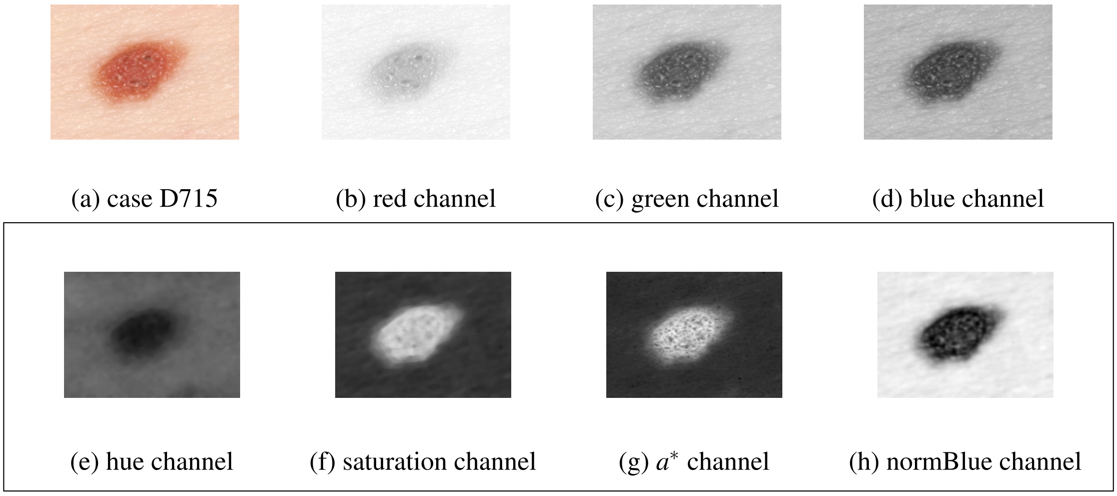

(a) case D364
(b) red channel
(c) green channel
(d) blue channel
(e) hue channel
(f) saturation channel
(g) a^* channel
(h) normBlue channel

Figure 4.1: Colour properties for non-pigmented lesion case D364. (b)-(d): properties directly from RGB colour channel. (e)-(h): properties extracted and selected in our work



(a) case D715 (b) red channel (c) green channel (d) blue channel

(e) hue channel (f) saturation channel (g) a^* channel (h) normBlue channel

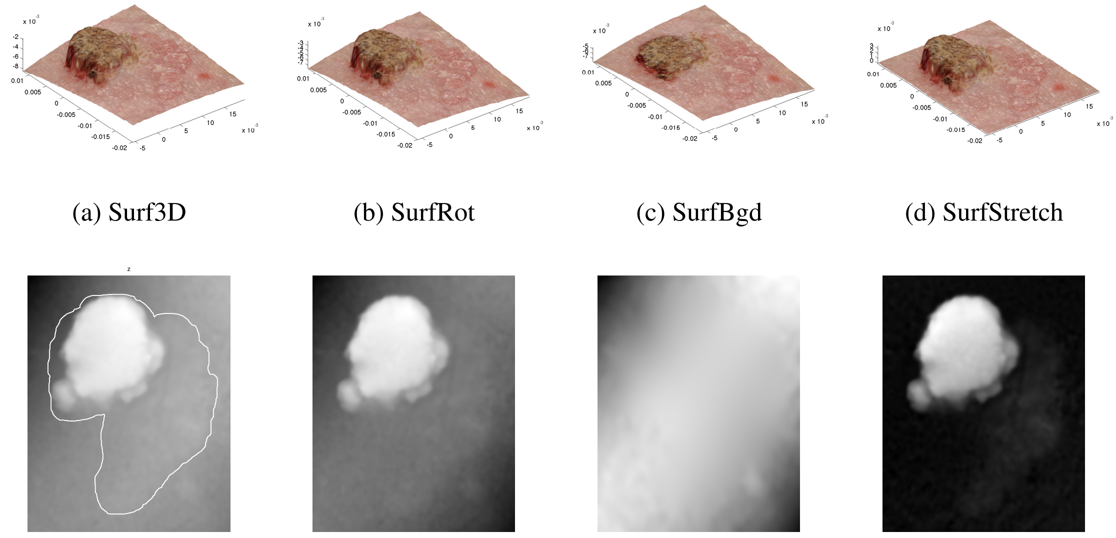
Figure 4.2: Colour properties for pigmented lesion case D715. (b)-(d): properties directly from RGB colour channel. (e)-(h): properties extracted and selected in our work

Relative depth We hypothesise geometric attributes should provide some complementary information to assist the segmentation. The question arises as how to extract depth feature from the data.

As our stereo imaging system obtains the superficial geometric information in terms of a 3D point cloud, the resulting data cannot be directly used in our segmentation model. In order to make them applicable, some specific processes are necessary:

1. Fit the lesion surface with the 3D point cloud as *Surf3D*. The surfaces are shown in *Fig.* 4.3(a) and 4.4(a).

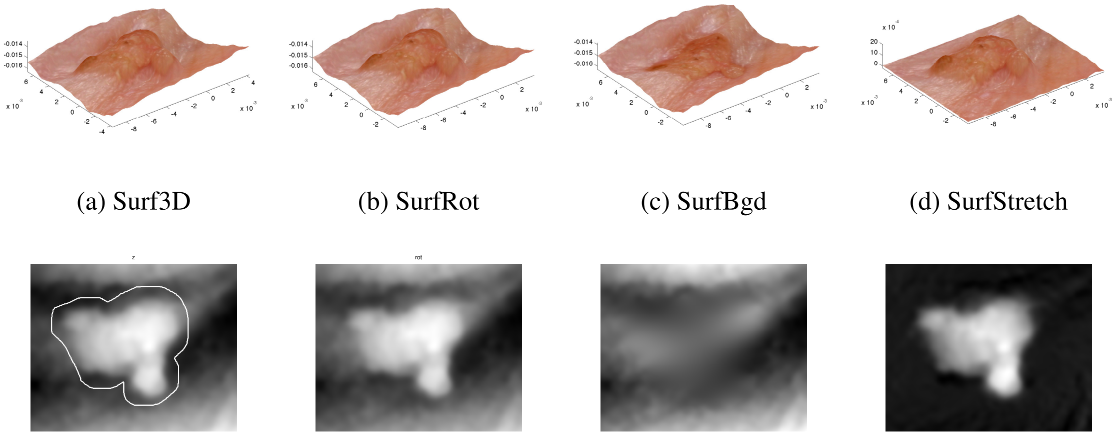
2. Rotate the lesion surface so that it would face the viewer. This involves 1) finding the 3D coordinate basis of the lesion surface (3 orthonormal vectors denoted as $x_{lesion}, y_{lesion}, z_{lesion}$) and 2) performing transformations so as to align the 3D lesion coordinate axis with the 3D coordinate axis associated with the imaging system (x, y, z), in which the z-axis faces the viewer. The 3D coordinate basis of the lesion surface at a point (*e.g.,* the center point of the surface, denoted as O) is derived using the global normal of the surface, which is estimated as the normal to a plane fitted using the 3D points of the lesion surface. This normal is thereby considered as the z-axis of the lesion surface (z_{lesion}). The x-axis (x_{lesion}) is taken as a random vector in the fitted 3D plane which passes O. The cross product of the z-axis with x-axis defines the y-axis (y_{lesion}). The rotation matrix can be



(a) Surf3D (b) SurfRot (c) SurfBgd (d) SurfStretch

(e) Original 3D data (f) Rotated 3D data (g) Body curvature surface (h) Flattened 3D data

Figure 4.3: 3D data transformation of the non-pigmented lesion case D364. (a), (b), (c) and (d): colour textured 3D model at each transformation step. (e), (f), (g) and (h): corresponding depth image



(a) Surf3D (b) SurfRot (c) SurfBgd (d) SurfStretch

(e) Original 3D data (f) Rotated 3D data (g) Body curvature surface (h) Flattened 3D data

Figure 4.4: 3D data transformation of the non-pigmented lesion case D570. (a), (b), (c) and (d): colour textured 3D model at each transformation step. (e), (f), (g) and (h): corresponding depth image

thereby obtained as $R = [x\ y\ z] \times [x_{lesion}\ y_{lesion}\ z_{lesion}]^{-1}$. Hence, the x, y and z coordinates of the point cloud of the lesion surface can be rotated using R. The rotated surface is denoted as *SurfRot* and demonstrated in *Fig.* 4.3(b) and 4.4(b).

3. Flattening the surface to remove the influence of the body curvature. The 'flatten' is achieved by a differencing operation between the rotated surface *SurfRot* and a background surface that denotes as *SurfBgd*. *SurfBgd* accounts for the local surface shape (shown in *Fig.* 4.3(c) and 4.4(c)). It is fitted as a quadric surface using only the 3D points on the normal skin region (This skin region lies outside the dilation of the initial contour (manually outlined) which is displayed in *Fig.* 4.3e and 4.4e in white). Mathematically, the resulting flattened surface *SurfStretch* = *SurfRot* - *SurfBgd*. Examples of the flattened surfaces are shown in *Fig.* 4.3(d) and 4.4(d).

From the above figures, it can be seen that this pre-processing of the 3D data is necessary in the light of highlighting the shape variations. If we project the *SurfStretch* onto the x-y axis plane, it results in a depth image (*e.g., Fig.* 4.3(h) and 4.4(h)) which represents the height information at each lesion pixel. After applying spatial filtering using the adaptive anisotropic diffusion, one can obtain a smoother depth image, which can be directly used in the segmentation. We refer to this as the depth-based feature and denote it as D. It is worth emphasizing that D is selected in the 3^{rd} place in the greedy feature selection procedure when we add it as an additional modality to the property pool. This supports our claim that geometric property does provide extra information for lesion separation and improves the segmentation result.

Texture

Currently, local texture descriptor based on the gradient structure tensor has a good reputation because of its good properties for texture discrimination [86]. Therefore, we choose it as our textural feature. The gradient structure tensor is a matrix of first partial derivatives and has the form as $J = \begin{pmatrix} I_{x1}^2 & I_{x1}I_{x2} \\ I_{x1}I_{x2} & I_{x2}^2 \end{pmatrix}$. The matrix J yields three different texture properties at each image location x. Often, the matrix J is replaced by its square root. This normalization step ensures that all feature channels have approximately the same dynamic range and generates three texture properties as [92]

$$f(x) = (J_1, J_2, J_3) = \left(\frac{I_{x1}^2}{|\nabla I|}, \frac{2I_{x1}I_{x2}}{|\nabla I|}, \frac{I_{x2}^2}{|\nabla I|} \right). \tag{4.34}$$

The first derivatives (I_{x1} and I_{x2}) of an image are not rotationally invariant. To compensate, we adopt the steerable Gaussian filter proposed in [95] to calculate the directional derivative I_{x1} oriented at angle α with respect to the x-axis and I_{x2} at degree $\alpha + 90^o$. α starts at 0^o degrees and increases by 15^o until 75^o. The texture property at each image location is the average. Hence, $I_{x1} = \overline{\{I_{x1}(\alpha), \alpha = 0, 15, \ldots, 75\}}$ and $I_{x2} = \overline{\{I_{x2}(\alpha + 90), \alpha = 0, 15, \ldots, 75\}}$. Next, we condense the texture feature by
1) collecting only the maximum tensor of the individual property channels (*i.e.,* the four colour and one depth). For example, $J_1 = I_{x1}^2 = \max\{\frac{I_{ix1}^2}{|\nabla I_i|}, i = \{1, \ldots, M\}\}$, M is the number of colour and depth features. Here, M = 5.
2) collecting only the maximum tensor of the three channels (*i.e.,* J_1, J_2, J_3). This is in the interest of reducing the feature space redundancy and improving the rotation invariance property. Thus, we have

$$f_{ST}(x) = \max\{J_1, J_2, J_3\}. \tag{4.35}$$

Furthermore, for the structure tensor based features, a nonlinear edge preserving smoothing process is often coupled in order to deal with noise and outliers in the data and ease further processing. Thereby, an adaptive anisotropic diffusion method is applied to smooth homogeneous regions while inhibiting diffusion in highly textured regions. For a given texture property image (*e.g.,* $I = f_{ST}$), the filtering process is implemented as:

$$\frac{\partial I}{\partial t} = \nabla \cdot [c(|\nabla I_\sigma|) \nabla I], \qquad I(t = 0) = f_{ST}. \tag{4.36}$$

The diffusion conductance c is image dependent and varies as a function of the derivative of the image I_σ (the Gaussian-smoothed version of I) at time t. In order to control the diffusion near the edges, Perona and Malik [96] defined $c(x) = \exp\left(-\frac{x^2}{P^2}\right)$, where the constant P was determined empirically. The function value is small where the gradient of the property image is large, resulting in lower diffusion near the textured locations like boundaries [97]. In other words, diffusion across the textures can be prevented while allowing diffusion along the texture. Hence, this anisotropic diffusion prevents the edge from being smoothed during the filtering process. Two modifications are applied to improve the performance of the diffusion filter. First, the property image is smoothed by a Gaussian filter with parameter σ decreasing at each iterations rather than a fixed σ value. In Perona and Malik's work [96], the reason that they calculated the gradient of the image based on the smoothed version I_σ rather than directly on I is to solve the ill-posed problem (the gradient measurements on I is not reliable because of noise, *e.g.,* where images close to each other could produce divergent solutions and very different edges). However, with the image (I) is getting smooth, the non-informative noise diminish faster than useful edge information, the gradient measurements calculated on I become more reliable and more informative. To account for this, the Gaussian smoothing parameter σ should evolve (decrease) with the time (t) as $\sigma = \sigma_0 \times (1 - 1/(T+1))^t$, where $\sigma_0 = 0.5$ and T denotes the time duration for the evolution of the diffusion function and it is determined experimentally as 20 iterations. Second, P is computed adaptively as a function of time - large in the beginning and get smaller gradually (*i.e.*, $P = P_0 \times (1 - 1/(T+1))^t, P_0 = 1$). This allows the noise to be reduced significantly at the beginning of the filtering process and the edges with different level of gradient to be enhanced at different times in the evolution [97]. The diffused structure tensor images are given in *Fig.* 4.5c and 4.6c. The textural difference can be seen between the lesion and its surrounding skin.

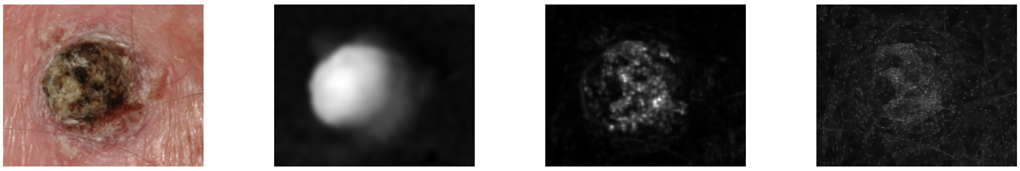

(a) colour (b) depth (c) structure tensor (d) texture scale

Figure 4.5: Texture properties for non-pigmented lesion case P299

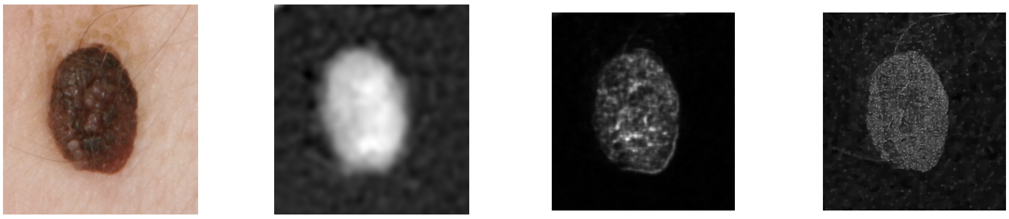

(a) colour (b) depth (c) structure tensor (d) texture scale

Figure 4.6: Texture properties for non-pigmented lesion case D374

However, the structure tensor based features only hold the information of the orientation and magnitude of a texture. The local scale information is missing. Brox *et al.* [92] pointed out that the scale was also an important aspect in discriminating lesion and healthy skin as textures were observable on different ranges of scale. In this context, we include a local scale feature to our property pool using the measurement technique introduced by Brox *et al.*. In their work, a region based local measure is calculated based on the assumption that pixels change their value with a speed that is inversely proportional to the size of the region they belong to. More details can be found in the paper [92]. Examples are shown in *Fig.* 4.5d and *Fig.* 4.6d. This feature is denoted as f_{scale}. Hereby, we have

$$I_{texture}(x) = \{f_{ST}(x), f_{scale}(x)\}. \tag{4.37}$$

We refer the texture-based feature as T.

For the purpose of quantitatively representing the local texture information of skin lesions, we also apply the Histogram of Oriented Gradient (HOG) descriptor. This descriptor captures the local coherence of object appearance. HOG was first introduced by Dalal *et al.* in 2005 for detecting humans in static imagery [98]. Since then, HOG has been extensively used in computer vision and image processing for object detection and segmentation. It has become the state of the art in these tasks [99]. The essential idea behind the Histogram of Oriented Gradient descriptors is that local object appearance and shape within an image can be described by the distribution of intensity gradients or edge directions [99]. Thereby, the implementation of HOG is based on the technique that counts occurrences of gradient orientation in localized portions of an image. The details can be summarized in the following steps: 1) divide the image into small connected regions, referred to as cells, 2) for each cell, compile a histogram of gradient directions or edge orientations for the pixels within the cell and 3) normalize and combine these histograms to represent the descriptor. For extending the HOG descriptor to multiple channel data, one can compute the gradient in each of the channels (colour or depth), respectively, and then select the response with greatest magnitude. Full detailed description of the HOG can be found in [98]. As HOG can derive a dense grid of uniformly spaced cells, we adapt the HOG descriptors by defining the cell as the neighborhood region of each image location so that each pixel corresponds to a high dimensional histogram vector. We therefore treat it as a single texture feature and denote it as f_{HOG}.

Hence, the texture feature vector at each pixel x can be extended to

$$I_{texture}(x) = \{f_{ST}(x), f_{scale}(x), f_{HOG}(x)\}. \tag{4.38}$$

We have now obtained a property vector with colour (C), depth (D) and texture based properties (T). So far, these three categories of features are only defined by the available image sensors and the intuitive feeling based on the visual observation, but not by the requirements of an optimal segmentation. An intensive analysis of the contributions of different characteristics, particularly focused on the impact on the 3D-based lesion information will be demonstrated through multiple experiments in the following Section 4.3.4. First, we need to consider how to model these features statistically.

4.3.3 Distribution Models

The features in different regions have different underlying distributions. They can be modeled by different Probability Density Functions (pdf). Deng *et al.* [79] addressed the fact that region based segmentations were highly dependent on these regional models. Generally, the $pdfs$ can be represented parametrically or non-parametrically. We prefer the former because it is more mathematically tractable and has been popular in the field. Hence, the $pdfs$ $p_i(I|\Omega_i)$ in Ω_i can also be represented as $p_i(I|\theta_i)$. For a particular choice of parametric density model, θ is estimated from the feature space of the associated regions and updates with the evolution of the contour.

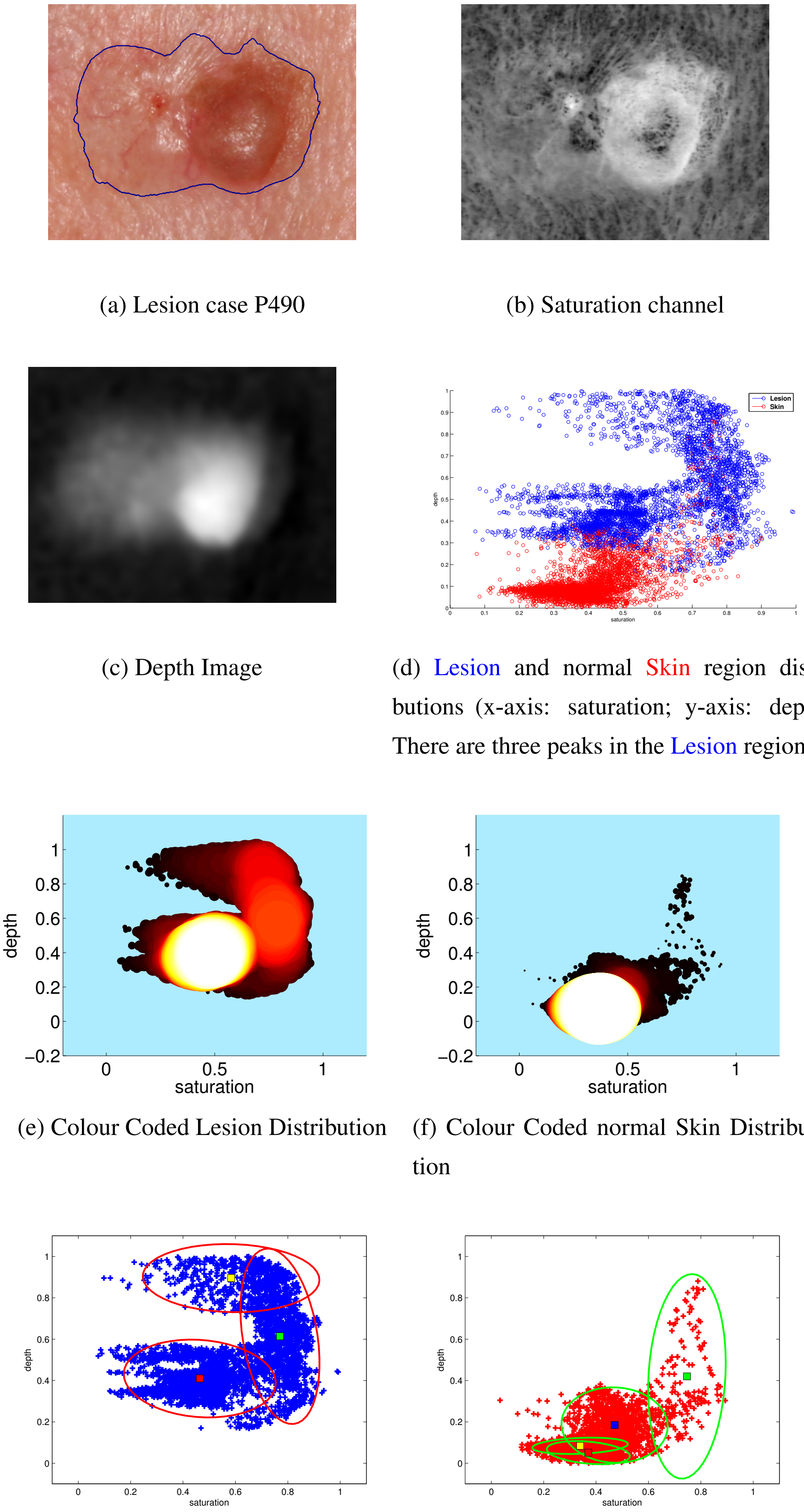
In the following, we will introduce different parametric density models, from which we can choose the best one for our application.

1. **Multivariate Gaussian Mixture Model (GMM)**

   Visual inspection reveals that the lesion region usually has inhomogeneous contents due to inherent nature, especially for BCC and SCC (this is also the case for the background normal skin region because of hairs and skin markings). Hence, the distribution of a property often has multiple peaks as shown in *Fig.* 4.7. For display purposes, we only show the two-dimensional histogram based on the saturation and depth feature values. This inhomogeneity can be resolved using a Gaussian mixture model (GMM). In *Fig.* 4.7g and 4.7h, the lesion region is modeled using a three component GMM; while the skin region is modeled using a four component GMM.

   The parameters of a GMM ($\theta_i = \mu_{ij}, \Sigma_{ij}, p_{ij}, j = 1, \dots, K_i$) are estimated using the approach proposed by Ma *et al.* [100], who used a dynamic merge or split learning strategy to determine the mixture component number and other corresponding parameters adaptively.

   Comparatively, the GMM seems to be the best option to model the density dis-



(a) Lesion case P490

(b) Saturation channel

(c) Depth Image

(d) Lesion and normal Skin region distributions (x-axis: saturation; y-axis: depth). There are three peaks in the Lesion region

(e) Colour Coded Lesion Distribution

(f) Colour Coded normal Skin Distribution

(g) Lesion region GMM

(h) Skin region GMM

Figure 4.7: The bivariate density distribution of the lesion and skin regions using the saturation (b) and depth (c) channels

tribution of both lesion and skin regions. However, because of its heavy calculation, some simplifications are considered.

2. **Multivariate Gaussian Model (MGM)**

   In practice, even though the distribution of data is not a Gaussian, the Gaussian density can still be used to approximate it, since a unimodal distribution is expected [79]. Therefore, we assume that the feature space follows a multivariate Gaussian distribution. The parameters of the model, a vector mean and a covariance matrix are determined as the maximum-likelihood estimators from the observations in individual regions:

$$\mu_i = \frac{\int_{\Omega_i} I(x)dx}{\int_{\Omega_i} dx} \tag{4.39}$$

$$\Sigma_i = \frac{\int_{\Omega_i} (I(x)-\mu_i)(I(x)-\mu_i)^T dx}{\int_{\Omega_i} dx}. \tag{4.40}$$

   Hence, the PDE in *Fig.* 4.13 can be extended as

$$\begin{aligned} \frac{\partial\phi(x)}{\partial t} &= -\frac{\partial E_1(\phi)}{\partial\phi} && (4.41)\\ &= \delta(\phi(x))\left(\log\frac{p_1(I(x)|\Omega_1)}{p_2(I(x)|\Omega_2)}\right) && (4.42)\\ &= \delta(\phi(x))\left(\log\frac{p_1(I(x)|\mu_1,\Sigma_1)}{p_2(I(x)|\mu_2,\Sigma_2)}\right) && (4.43)\\ &= \delta(\phi(x))\log\frac{|\Sigma_1|^{-1/2}\exp\left(-\frac{1}{2}(I(x)-\mu_1)^T\Sigma_1^{-1}(I(x)-\mu_1)\right)}{|\Sigma_2|^{-1/2}\exp\left(-\frac{1}{2}(I(x)-\mu_2)^T\Sigma_2^{-1}(I(x)-\mu_2)\right)} && (4.44) \end{aligned}$$

   From the pattern recognition point of view, this can be considered as a Quadratic Discriminant Analysis (QDA) between two classes (*i.e.,* lesion and skin). The separating surface is quadratic. The likelihood ratio associated with the regional competition result between the two classes leads to the force that drives the level-set evolution.

3. **Independent Feature Model (IFM)**

   When the off-diagonal elements of the covariance matrix (Σ) in MGM are forced to be 0, the features are considered as uncorrelated and independent. The multivariate density distribution are further simplified as a *naive* Bayes probability model:

$$p_i(I(x)|\mu_i,\widehat{\Sigma_i}) = \frac{\exp\left(-\frac{1}{2}(I(x)-\mu_i)^T\widehat{\Sigma_i}^{-1}(I(x)-\mu_i)\right)}{(2\pi)^{\frac{dim}{2}}|\widehat{\Sigma_i}|^{\frac{1}{2}}} \tag{4.45}$$

$$= \prod_{j=1}^{dim} \frac{\exp\left(-\frac{1}{2}(f_j(x)-\mu_{ij})^2/\sigma_{ij}^2\right)}{\sqrt{2\pi\sigma_{ij}^2}} \tag{4.46}$$

$$= \prod_{j=1}^{dim} p_i(f_j(x)|\mu_{ij},\sigma_{ij}), \tag{4.47}$$

where, $p_i(f_j(x)|\mu_{ij},\sigma_{ij})$ describes the j^{th} feature in a region Ω_i and

$$\widehat{\Sigma}_i^{-1} = \begin{vmatrix} \frac{1}{\sigma_{i1}^2} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sigma_{i2}^2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sigma_{idim}^2} \end{vmatrix}.$$

In spite of the fact that this independence assumption is too strong and is often invalid in practice, this density model has the advantages of simplicity and low sensitivity to not-informative variables (*i.e.,* noise) [101]. Therefore, it has worked quite well in many complex real-world situations [79], *e.g.,* the naive Bayesian Classifier in the pattern recognition field. In our application, the extracted features are associated with different types of lesion characteristic and are properly selected (for the four colour features). We can therefore assume that they have low correlation and fulfill the independent requirement and can be modeled in this way.

4. **Mumford-Shah Model (MSM)**

If we further strengthen the assumption by forcing all the features to carry the same variance $\sigma = \sigma_{ij}, \ \forall\, i,\ j$ in the covariance matrix

$$\widehat{\Sigma}_i^{-1} = \begin{vmatrix} \frac{1}{\sigma^2} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sigma^2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sigma^2} \end{vmatrix},$$

then we assume they are standardized and uncorrelated. The probability model of region Ω_i becomes:

$$p_i(I(x)|\mu_i,\widehat{\Sigma}_i) = \prod_{j=1}^{dim} \frac{\exp\left(-\frac{1}{2}(f_j(x)-\mu_{ij})^2/\sigma^2\right)}{\sqrt{2\pi\sigma^2}} \tag{4.48}$$

$$= \prod_{j=1}^{dim} p_i(f_j(x)|\mu_{ij},\sigma). \tag{4.49}$$

This leads to a PDE as

$$
\begin{aligned}
\frac{\partial \phi(x)}{\partial t} &= -\frac{\partial E_1(\phi)}{\partial \phi} && (4.50)\\
&= \delta(\phi(x))\left(\log \frac{p_1(I(x)|\Omega_1)}{p_2(I(x)|\Omega_2)}\right) && (4.51)\\
&= \delta(\phi(x))\left(\log \frac{p_1(I(x)|\mu_1,\sigma)}{p_2(I(x)|\mu_2,\sigma)}\right) && (4.52)\\
&= \delta(\phi(x))\log \prod_{j=1}^{dim} \frac{\exp(-(f_j(x)-\mu_{1j})^2/2\sigma^2)}{\exp(-(f_j(x)-\mu_{2j})^2/2\sigma^2)} && (4.53)\\
&= \delta(\phi(x))\log \prod_{j=1}^{dim} \exp \frac{-(f_j(x)-\mu_{1j})^2+(f_j(x)-\mu_{2j})^2}{2\sigma^2} && (4.54)\\
&= -\frac{1}{2\sigma^2}\delta(\phi(x))\sum_{j=1}^{dim}[(f_j(x)-\mu_{1j})^2-(f_j(x)-\mu_{2j})^2] && (4.55)\\
&= \frac{1}{2\sigma^2}\times\delta(\phi(x))[(I(x)-\mu_2)^2-(I(x)-\mu_1)^2]. && (4.56)
\end{aligned}
$$

As $\frac{1}{2\sigma^2}$ is only a constant ratio in each iteration, it can be ignored. It turns out that *Eq.* 4.56 is equivalent to the PDE used in the Chan-Vese algorithm which is derived from the Mumford-Shah functional [85]. The smooth approximations of individual regions are in fact μ_1 and μ_2. From another point of view, the Chan-Vese algorithm is the most simplified version in the statistical regional-based segmentation. As mentioned before, this approach has been very commonly used in the field. It works well in the situations where the features are piece-wise constant, which is obviously not true in our application. We will therefore use it in our experiment as a baseline model.

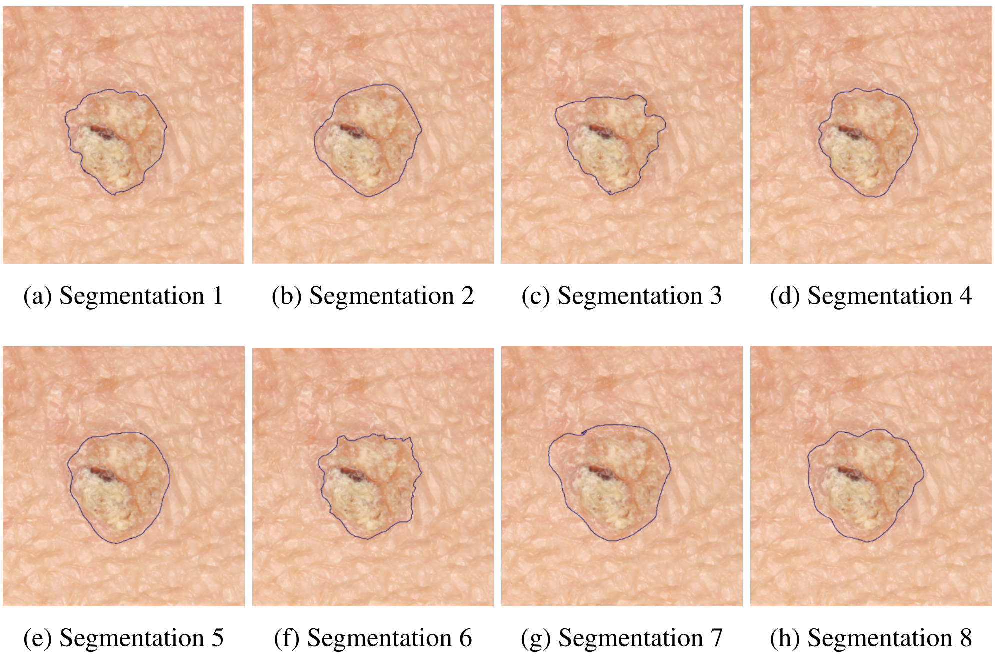
4.3.4 Experiments and Results

The 50 test images used in our comparison are randomly selected from our lesion database including five different classes of lesions: **SCC (Squamous cell carcinoma)**, **ML (Melanocytic nevus)**, **BCC (Basal cell carcinoma)**, **AK (Actinic Keratosis)** and **SK (Seborrhoeic Keratosis)**. 21 of them are pigmented lesions[2]; while the other 29 are non-pigmented. The separation of pigmented and non-pigmented lesions is done by Dr. Ben Aldridge. Pigmented lesions have browny colour and non-pigmented

[2] There are two reasons for lesions being pigmented. It is either due to melanocyte hyper proliferation, such as in a melanoma or a melanocytic nevus or, alternatively, as in a seborrheic keratosis when the number of melanocytes is normal but they seem to produce too much melanin.

lesions are light-coloured. The manual segmentations of these lesions were given by eight dermatologists from the Dermatology department of Edinburgh University and are used for performance evaluation. A Standard Tumour Area (**STA**) is defined as the region determined using our ground truth estimation method, **LSMLP**, which is proposed in Chapter 3. An extra manual segmentation by a 9^{th} dermatologist is used for comparing against the computer-based results.

The standard deviation (**SD**) of lesion areas derived from the eight manual segmentations is calculated for each lesion image. To make them comparable over 50 test images, the value is normalized by the corresponding **STA**. The average **SD** over our 50 test images is 14.26%. This result confirms the conclusion in Chapter 3 that there is large variation between manual segmentations. Further analysis also shows that there is more variation in clinical opinion of lesion boundaries for non-pigmented lesions (**SD** = 18.28%) than pigmented ones (**SD** = 8.71%). An example of the human segmentations of a non-pigmented lesion is shown in *Fig.* 4.8. This indicates the difficulty for non-pigmented lesion segmentations.


(a) Segmentation 1 (b) Segmentation 2 (c) Segmentation 3 (d) Segmentation 4

(e) Segmentation 5 (f) Segmentation 6 (g) Segmentation 7 (h) Segmentation 8

Figure 4.8: Manual segmentations of case D647

For each lesion data, the uniform segmentation method described in Section 4.3 is performed four times using different property combinations as follows [3]:

[3] In this thesis, we only compare the following four feature combinations in which the colour plays

colour (C):

$I_C = (f_{hue}, f_{saturation}, f_{a*}, f_{blue})^T$.

colour + depth $(C+D)$:

$I_{CD} = (f_{hue}, f_{saturation}, f_{a*}, f_{blue}, f_{depth})^T$.

colour + depth + texture $(C+D+T)$:

$I_{CDT} = \left(f_{hue}, f_{saturation}, f_{a*}, f_{blue}, f_{depth}, f_{ST}, f_{scale}\right)$.

colour + depth + HOG $(C+D+HOG)$:

$I_{CDT(hog)} = \left(f_{hue}, f_{saturation}, f_{a*}, f_{blue}, f_{depth}, f_{HOG}\right)$ [4].

For each property combination, the distribution models (detailed in Section 4.3.3) are fitted on the data. It is worth noting that the original colour-based feature set is narrowed down to four components using a greedy property selection method. The selection is based on the segmentation error rate criterion. The number of features is determined by the change of the error criterion (*e.g., XOR*), which decreases at the beginning and increases after the 4th feature is added (see *Fig.* 4.9). This strategy on one hand reduces the redundancy between features and on the other hand, keeps only the most representative and informative features for segmentation. As we consider this property selection step as a parameter tuning process, we do not split the data into training and testing sets. Nevertheless, the feature selection result is found consistent because the feature set selected using the 50 data set (details can be found in Section 4.3.4) and 20 data set (our previous segmentation data set used in [22]) are identical.

In addition, for **deformable model** based segmentations, an initial contour is always needed. Good initial contour location is helpful for the statistical regional based approaches as the final result is sensitive to the initial setting. Normally, the contour is drawn manually by users, therefore, most methods are semi-automatic. In order to achieve fully autonomy, we perform a clustering based approach to pre-segment the lesion in order to obtain the initial contour. Because different regions of the lesion are represented using the selected properties, which, in turn, have different statistical distributions, we roughly model them as two components of a Gaussian mixture model using the Expectation Maximization (**EM**) algorithm. By comparing the density val-

the dominant role because 1) colour has been proved as the most important information for lesion analysis [36] and should not be neglected in the segmentation task, 2) the aim of the chapter is to address whether the depth data will further improve the lesion segmentation when it is integrated with the colour data (in other words, to demonstrate whether the segmentation using $C+D$ will exceed C) and 3) the depth or the texture or the combination of the two fail on certain lesion data which are flat and smooth (*e.g.,* junctional nevus)

[4] For the HOG based features, instead of putting it together with the structure tensor and scale based texture properties, we treat it as an independent texture descriptor because we would like to evaluate its capability to represent the texture information of skin lesions

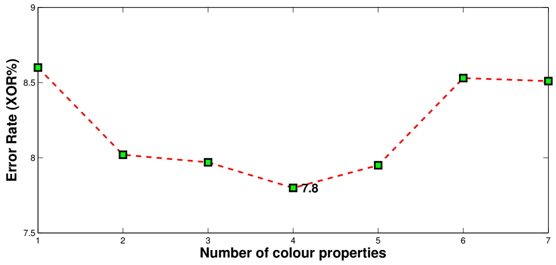
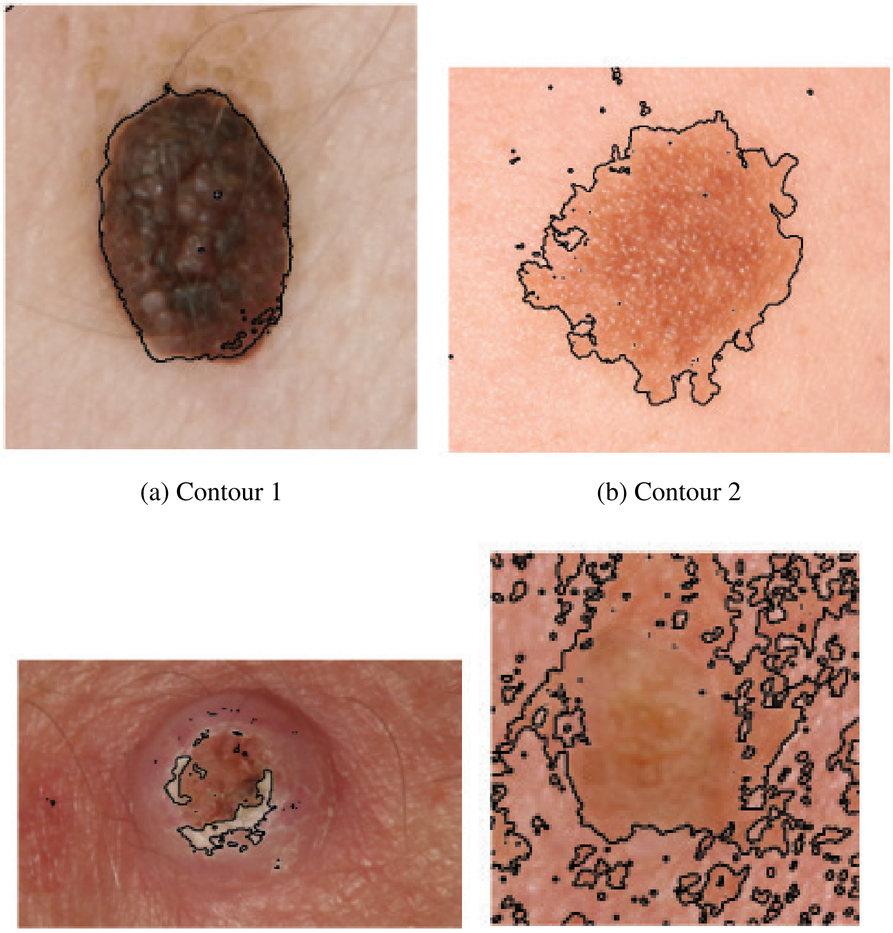


Figure 4.9: Greedy colour property selection. The segmentation error rate decreases at the beginning with the additional property selected using the greedy strategy and increases when more than four properties are used

ues of each pixel to the two Gaussian components, the lesion image is labeled as a binary image. The lesion region (foreground) can be defined as the region with less position (x, y coordinate) deviation considering the lesion pixels are normally grouped at the center of the data. Two examples are shown in *Fig.* 4.10 (a) and (b). However, this strategy may fail in a few cases (*e.g.,*84 out of 867 lesion data cannot obtain a correct initial segmentation using this method. Some examples of failures as shown in *Fig.* 4.10 (c) and (d)), because of the complexity of lesions: 1) artifacts or intrinsic cutaneous features like hairs and bloody vessels, 2) the nature of lesions (*e.g.,* non-pigmented, bad growing positions). Human interaction is necessary in those cases.

To evaluate computer-based segmentations, a quantitative metric is important. Some recent works have proposed to use the (Normalized) Probabilistic Rand Index - (N)PRI. The (N)PRI evaluates the divergence of a segmentation (S) from the 'ground truth' (GT) by measuring the fraction of GT that agree with S on labeling a pair of distinct pixels differently or identically [64, 70]. However, after several cases study, Peserico *et al.* [102] concluded that (N)PRI suffered from certain shortcomings (due to its non-monotonicity with the fraction of misclassified pixels) and raised doubts on adopting the (N)PRI over the simpler and established metrics XOR (or Error rate). They concluded that XOR was preferable for the evaluation of lesion segmentations. Thereby, XOR is chosen as our evaluation tool in the following analysis and has the format as $XOR = \frac{Area(\mathbf{GT} \oplus \mathbf{SEG})}{Area(\mathbf{GT} + \mathbf{SEG})}$ (**GT** is the ground truth and **SEG** denotes the computer-based segmentation result). Based on it, we performed a comparison study of the segmentations using four different combinations of properties as follows:

The segmentation results of the feature combinations are shown in *Table.* 4.1. Both



(a) Contour 1

(b) Contour 2

(c) Contour 3

(d) Contour 4

Figure 4.10: Initial Contour. (a)-(b): successful examples; (c)-(d): failed examples

the error rate and the variation are given. It shows that:

1. Based on the overall error rate, the worst uniform segmentation result is obtained when using only C. The feature set that integrates colour and depth information outperforms colour alone and improves the error rate from 7.80% to 6.78%. This convincing gain suggests that the depth descriptor is complementary. Otherwise, overall performance would not have improved significantly. It is also worth mentioning that although we only did forward feature selection on the colour features, when we add depth to the pool, it is also selected as a valuable feature. This reflects that the segmentation of lesions benefit from the additional depth property.

2. The best uniform segmentation result (average error rate of 6.62%) comes from the feature set that integrates all the properties derived from colour, depth and

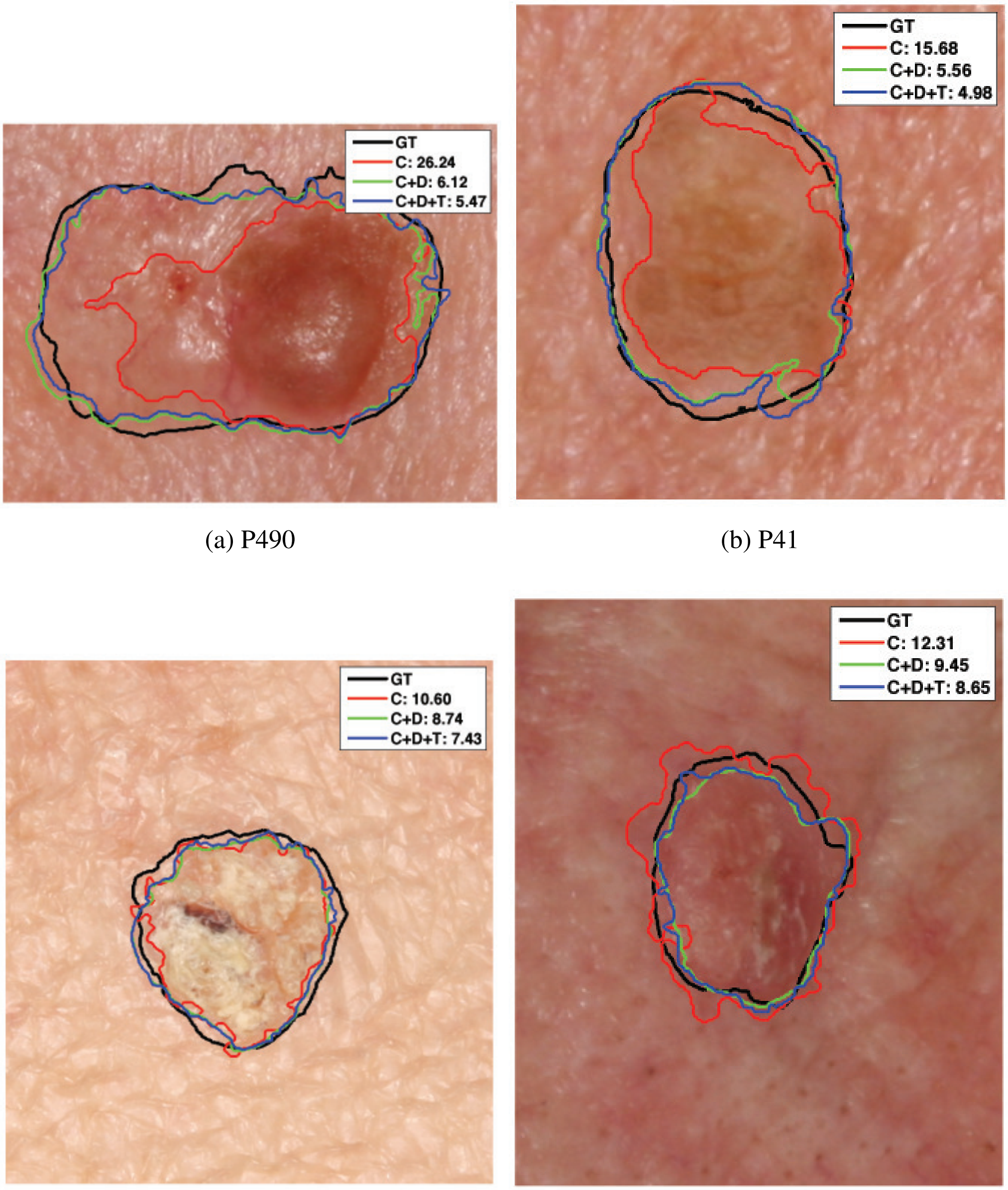
Error Rate %(XOR)	Overall	Pigmented	Non-pigmented
Uniform method (C)	7.80(±5.35)	5.34(±3.18)	9.59(±5.91)
Uniform method (C+D)	6.78(±3.05)	5.87(±2.43)	7.44(±3.32)
Uniform method (C+D+T)	6.62(±2.60)	5.89(±2.46)	7.15(±2.61)
Uniform method (C+D+HOG)	7.59(±3.23)	6.21(±2.50)	8.58(±3.41)
9^{th}Dermatologist	7.02(±5.23)	5.30(±3.60)	8.08(±5.61)
Lowest Error Rate Combination	5.66(±2.48)	5.25(±2.09)	5.95(±2.97)
Uniform method (GMM)	6.62(±2.60)	5.89(±2.46)	7.15(±2.61)
Uniform method (MGM)	7.26(±3.84)	6.09(±2.78)	8.11(±4.29)
Uniform method (IFM)	7.53(±4.38)	5.87(±2.37)	8.72(±5.09)
Uniform method (MSM)	13.25(±3.88)	8.40(±5.49)	16.77(±9.22)

Table 4.1: Average segmentation error rates and their standard deviations. See text for discussion of different values

texture information. It produces the lowest error rate as 6.62%. However, when representing the texture characteristics using **HOG**, the result is not ideal. The failure in the HOG property might because of 1) the inappropriate selection of HOG scale, 2) the regional model for the HOG is not suitable and 3) it is also possible that HOG does not even fit the lesion segmentation application. However, $C+D+HOG$ still outperforms the feature set using only C. Therefore, we have a reason to believe that the texture also provides complementary information for segmentation. *Fig.* 4.11 shows contours obtained using different segmentation methods of several non-pigmented lesions. The error rate improves significantly by the integration of depth and texture information.

3. The dermatologist performs the best in segmenting the pigmented lesions, but there are large variations in clinical opinion of lesion boundaries for non-pigmented lesions, on which both $C+D$ and $C+D+T$ based segmentations perform better. In addition, the computer-based methods give more consistent results according to the lower standard deviation on both pigmented and non-pigmented lesions. This is a desired property for medical applications.

4. If we choose for each lesion the combination of properties with the lowest XOR measure, one could obtain an overall segmentation error rate as low as 5.66%. One can find that the choice of property combination is scattered and does not consist of a particular combination like $C+D+T$. Furthermore, if we consider the pigmented and non-pigmented lesions independently, one can find that the additional depth and texture information mainly makes contributions to the latter.



(a) P490

(b) P41

(c) D647

(d) D550

Figure 4.11: Integrating depth and texture information improves the segmentation results

The segmentations for non-pigmented lesions are significantly improved, but this is not the case for the pigmented lesions. Generally, the segmentation for the pigmented lesion using the depth data has a broader lesion region than that obtained by only using colour data. The lesion region is extended slightly to the adjacent region whose colour pigmentation has not been affected by the lesion. This is mainly because of the limited resolution of 3D data. Current 3D building techniques have not matured to a stage to capture all the details of the objects (a primary system accuracy test experiment is described in Appendix B). Certain levels of fine textural information like the adjacent region between the lesion and skin is blurred in the reconstruction step. However, this phenomenon might also be because of the development order of a lesion. For example, a lesion could first grow vertically and then exhibit pigment variation. For either reason, integrating depth information reduces the average error rate of the pigmented lesion segmentation based on the evaluation result. The lowest error rate for pigmented lesions are in fact produced by only using colour properties (*e.g.,* the segmentation for case P152 in *Fig.* 4.12). This reflects a conjunction between the lesion appearance (*e.g.,* pigmented or non-pigmented) and the choice of best properties for segmentation. In the next section, we will further explore this finding.

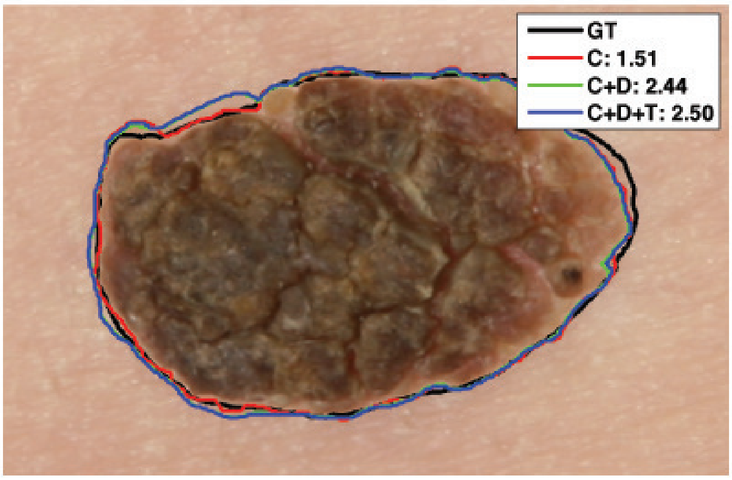


Figure 4.12: Integrating depth and texture information degrades the segmentation results for the lesion P152

5. We also compare the performances of different density models so as to determine the best statistical density model for representing the regional properties. The optional models include the **Multivariate Gaussian Mixture Model (GMM)**, the **Multivariate Gaussian Model (MGM)**, **Independent Feature Model (IFM)** and **Mumford-Shah Model (MSM)**. Their respective performances using $C+D+T$ features are listed in the bottom of *Table.* 4.1. It can be seen that the **GMM** provides the best result. For the purpose of reducing computational expenses, the **MSM** assumes that each region can be modeled

using a single Gaussian component, but this simplification reduces the model's capability of representing the complexity of real world lesions. The feature independent assumption in **IFM** further degrades the model performance. As proved before, the **MSM** is equivalent to the Chan-Vese algorithm [85] derived from the Mumford-Shah functional. In **MSM**, only the first order statistic (*i.e.,* mean) is taken into account, while all higher order statistics are ignored. **MSM** therefore over-simplifies the density distribution of lesion properties and does not suit the application. It achieves the worst result. We conclude that, for the complex components constitution of lesion data, the **GMM** is the optimal solution for modeling the regional characteristic based on both visual analysis and experimental results.

6. We conducted a statistical significance analysis of the segmentation results obtained using different properties based on the paired one-tailed t-test. The null hypothesis is that: **the error rate (XOR) provided by the $C+D+T$ and C based segmentations are the same. The alternative is that the former has a lower mean that the latter.** The resulting p-value on 50 trials is 0.0811. Under significance level 10%, the null hypothesis can be rejected, which means that the segmentation using $C+D+T$ provides improvement over that using C. Particularly, for the non-pigmented lesions, the improvements from C to both $C+D$ and $C+D+T$ are significant under a significance level 5%. The corresponding p-values are 0.0479 and 0.0231.

4.4 Decision Tree Based Segmentation

4.4.1 Method

In the uniform method, we treat all properties equally and simply stack them into a high-dimensional vector. As not all the derived properties are useful for every lesion, the irrelevant and redundant properties containing useless 'noise' could make the algorithms unstable and likely to converge to local minima. For example, even though the depth could provide extra information about the location of the lesion in general, there are cases for which the depth might be deemed as additional noise rather than useful information, *e.g.,* flat lesions. For such lesions, the colour and texture properties are good enough for segmentation. As a result, the question of how to pair different types of lesions with different combination of properties arises. As mentioned previously, there is a relationship between the lesion appearance (*e.g.*, pigmented or non-pigmented) and the choice of best properties for segmentation. We therefore combine pattern analysis techniques with the segmentation properties selection task. In this context, we propose a hierarchical strategy based segmentation algorithm.

4.4.2 Hierarchy

The decision-tree-based segmentation strategy aims to use different feature subsets based on an initial discrimination analysis between different lesion appearances. For a particular lesion, an automatic pre-categorization based on the lesion appearance is conducted according to a hierarchical tree structure. The tree leaf that it ends up shows the property combination (or subset) $f(x)$ that it should use for segmentation, which could be C or $C+D$ or $C+D+T$. The procedure is as follows:

1. As found in the previous section, for pigmented (brownly coloured) lesions, colour information can produce good segmentations, even better than integrating any other information. Therefore, in the first layer of the hierarchical structure (see *Fig.* 4.14a), colour is used to split different categories of lesions. For representing lesion colour information, we stick to the colour features used in lesion segmentation. The split criterion is based on the dissimilarity of regional colour values between the lesion and skin regions. The dissimilarity is calculated using the Euclidean distance between the mean colour value of the lesion and skin regions (*i.e.*, $|\overline{\{f_{hue}, f_{saturation}, f_{a*}, f_{blue}\}_{lesion}} - \overline{\{f_{hue}, f_{saturation}, f_{a*}, f_{blue}\}_{skin}}|$).

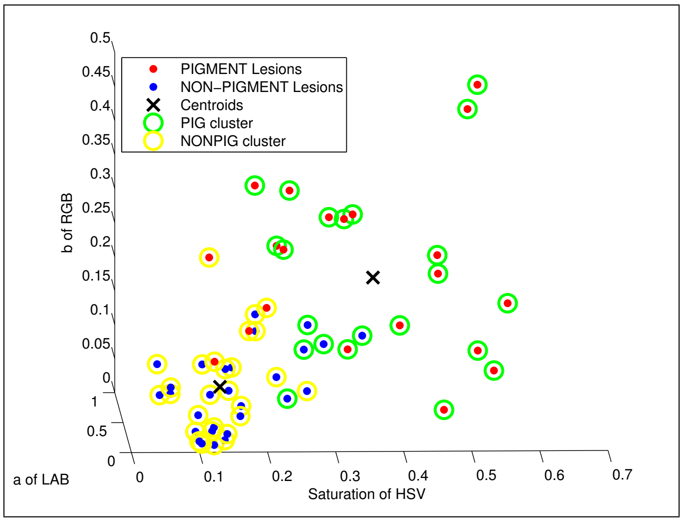


Figure 4.13: Pre-categorization of lesions based on colour properties

For a particular lesion, its dissimilarity value is compared to a pre-set threshold, based on which the lesion can be allocated to a corresponding category. This threshold is determined in a prior lesion pattern analysis and is equal to the midpoint between the 2 clusters found by using k-means (k=2) over the training data. *Fig.* 4.13 demonstrates the pattern analysis results. The two cluster centroids are highlighted with black crosses (X). Their corresponding components are highlighted using circles, where yellow is for lesions classified as non-pigmented and green for pigmented. The real category is also demonstrated. The pigmented lesions are shown as red spots and the non-pigmented ones as blue spots. According to *Fig.* 4.13, the clustering result and the truth are highly correlated. This indicates the goal of splitting the pigmented and non-pigmented lesions in the first layer of the hierarchical tree structure can be achieved.

2. The second layer uses depth information to split non-pigmented lesions into flat and non-flat categories (see *Fig.* 4.14b) based on the difference in depths between the lesion and normal skin. The threshold is chosen in a manner similar to the colour threshold.

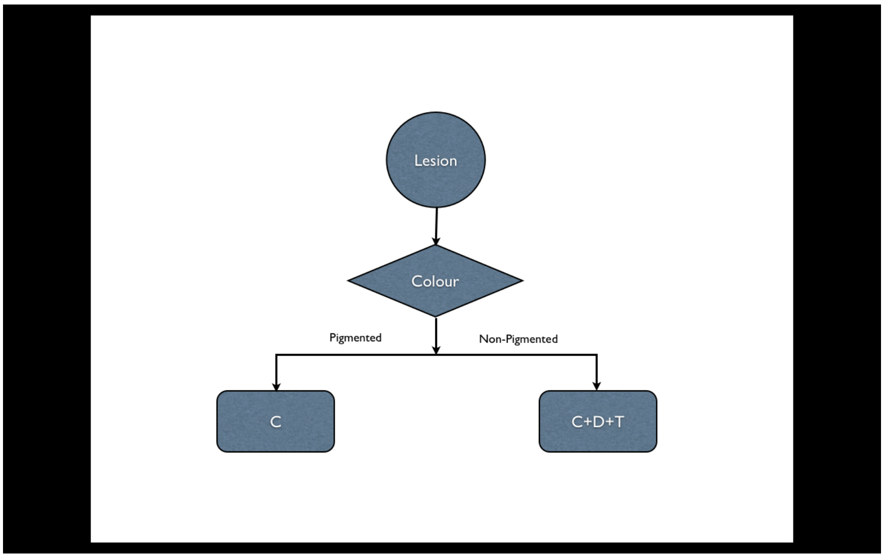
4.4.3 Experiments and Results

According to the hierarchical structure, the two splitting criteria could pre-categorize the lesion type, after which the corresponding optimal choice of feature properties should be applied for the segmentation. The results are summarized in *Table.* 4.2. It shows:

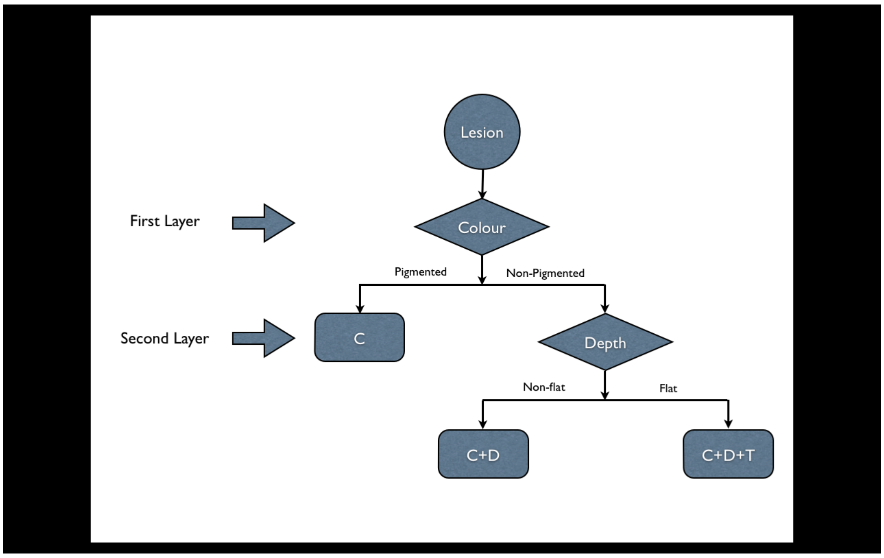
Error Rate %(XOR)	Overall	Pigmented	Non-pigmented
9^{th}Dermatologist	7.02(±5.23)	5.30(±3.60)	8.08(±5.61)
Uniform method (C+D+T)	6.62(±2.60)	5.89(±2.46)	7.15(±2.61)
Uniform method (PCA)	7.12(±3.96)	6.02(±3.41)	7.92(±4.19)
Hierarchical method (one layer)	6.56(±2.75)	5.03(±2.79)	7.05(±2.59)
Hierarchical method (two layers)	6.15(±2.58)	5.03(±2.79)	6.50(±2.44)

Table 4.2: Average segmentation error rates and their standard deviations

1. The one layer colour-based decision-tree-based segmentation reduces the error rate of the best uniform segmentation from 6.62% to 6.56% and the two layer colour and depth-based decision-tree-based segmentation achieved the lowest error rate of 6.15%. Though their overall improvements are not statistically



(a) One layer hierarchical structure



(b) Two layers hierarchical structure

Figure 4.14: Hierarchical structures

significant, the improvements on pigmented lesions are significant with the significance level of 5%.

2. To reduce the data dimensionality and find the most discriminative properties, Principal Analysis (PCA) is the most commonly considered solution. Thereby, we apply PCA to obtain the top three principal components from the original full property pool. The segmentation based on these three component is compared with the hierarchical strategy. It shows that the latter outperforms the PCA

approach which produced an error rate of 7.12%. In fact, the PCA based result is even worse than the best flat approach. This might be caused by the lost of discriminating information during the projection.

3. Pigmented and non-pigmented are two terms defined by dermatologists in order to separate different lesions. They do not influence the selection of features. They are only used here to evaluate the performances, *i.e.,* to tell the performances of segmentation methods on two lesion types (pigmented and non-pigmented). The reason I am doing this is to tell the impact of depth data on pigmented and non-pigmented lesions separately. Pigmented lesions have been widely studied in the field while non-pigmented lesions have been rarely discussed. In our work, we prove that the depth data is more meaningful for the non-pigmented lesions. In *Table.* 4.1 and *Table.* 4.2, features used for segmentation are selected using different strategies. In *Table.* 4.1, colour features are used for segmentation. In *Table.* 4.2, the features to be used for segmentation depend on the pre-category of the lesions. The pigmented lesions might be selected to the group that only uses colour features for segmentation, but they might also be automatically selected to other groups which use other features.

The pseudocode for both **Uniform segmentation** and **Decision-tree-based segmentation** is given in Appendix C.

4.5 Conclusion

Lesion segmentation is important as the classification rate depends highly on the accurate extraction of the lesion area. According to a study in [20] that evaluated the effectiveness of different lesion region extraction methods for the diagnostic accuracy, a good segmentation resulted in significantly improved diagnosis. However, obtaining a good segmentation result is a challenging task. The segmentation of non-pigmented lesions, which have been rarely considered in the literature but are included in our work, is especially difficult [52]. A good segmentation algorithm should produce segmentations with the characteristics of accuracy and consistency. In this chapter, we attempt to incorporate diverse image cues (*i.e.,* colour, depth and texture) to the segmentation model so that we can investigate whether or not the extra information would lead to better results. To build an experimental platform for comparison, we first presented a uniform segmentation algorithm. It is a region-based probabilistic formulation of the

deformable model that is implemented within the level-set framework. Colour, depth and texture features are measured at each image point. Their regional statistical parameters are estimated and used to characterize the respective lesion structures (*i.e.,* lesion and skin) with the density distribution model. We derive several density models and prove that the well-known Chan-Vese segmentation [85] is in fact equivalent to the most simplified version in the statistical regional-based segmentation. The **GMM** is selected as the best way to model the complex density distribution of properties upon comparison for our lesion segmentation application. In order to take into account the prior knowledge of spatial relationships, we further introduce the local spatial dependency term, which is modeled by MRF in the form of a Gibbs distribution.

The discriminative ability of various lesion information on separating lesion from skin is analyzed according to the segmentation accuracy metric **XOR**. Upon comparison, they all contribute to the region discrimination. Colour properties enable a close segmentation to the dermatologists on pigmented lesions. Integrating depth and texture properties results in an overall improvement of segmentation, in terms of both accuracy and consistency. The error rate is reduced from $7.80\% \pm 5.35\%$ to $6.62\% \pm 2.60\%$. This convincing gain suggests that the depth and texture descriptors are complementary. The depth and texture information is particularly meaningful for non-pigmented lesions as they have less colour variation over different regions. As a result, we have proved our claim that the depth information improves lesion segmentation.

As the analysis reveals that different information has specific importance for different types (*e.g.,* pigmented and non-pigmented) of lesions, we believe that only discriminative and useful properties should be chosen for each lesion according to their appearances, instead of using all the properties derived from all kinds of information for every lesion. Hence, we propose a novel hierarchal segmentation strategy. This segmentation approach is hierarchical in the sense that it uses different feature subsets within a hierarchical structure determined by the discrimination between different lesion appearances. The experiments show that both our one layer hierarchical approach based on colour and the two layer hierarchical approach based on colour and depth further improve the segmentation results compared to the uniform segmentation approach.

The novelties of our work are 1) we consider both pigmented and non-pigmented lesion data which increases the difficulty of our problem. The latter has rarely be considered in the literature as far as we can tell. 2) We comprehensively evaluate the contribution of various information on lesion segmentation, particularly focusing on the depth information based on the comparatively large and carefully validated lesion database. This has not been found in the literature. 3) We further incorporate pattern recognition techniques and propose a decision-tree-based segmentation structure which further improves the segmentation accuracy. 4) We prove the relationship between the segmentation model derived from the Mumford-Shah functional and the statistical regional-based segmentation model.

In the future, there are several potential improvements and follow-up work to be considered:

1. Our hierarchical structure is built up in an *ad-hoc* way according to some kinds of prior knowledge. In the future, a more sophisticated hierarchical structure can be built based on a better definition of the hierarchical layer splitting parameters.

2. On the other hand, in the process of finding the splitting parameter threshold, the segmentation data are not split into training and test sets. The test data themselves influence the value of the splitting threshold. This may result in that the threshold overfits the test data and loses universality especially when the database is of small size. A more proper way is to split data into training and test sets so that the lesion pre-categorization and the segmentation steps are separated. Additional lesion data with ground truth segmentations should allow more extensive development and testing.

Chapter 5

Classification

This chapter attempts to investigate the contribution of 3D information to lesion diagnosis through extensive and rigorous tests on both human and computer based classifications. First, Section 5.2 reviews each component in the lesion classification pipeline including feature extraction, feature selection and classifiers, followed by a prior study of the influence of 3D data in human diagnosis in Section 5.3. The investigation of the contribution of 3D data to a computer based five non-melanoma skin lesion classification task is presented in Section 5.4 based on a comparison between using colour features only and using both colour and depth features. Experiments and results in Section 5.5 show that adding the 3D-based features gives an improved classification rate compared to only using colour features.

5.1 Introduction

Segmentation allows computers to locate the problem skin region (*i.e.,* lesion). Classification enables computers to tell what the problem it is. As a high level data analysis procedure, classification is usually the final stage in the development of a computer aided diagnosis system. It often involves two steps: 1) representing lesion structures using a feature set that is extracted based on the segmentation results (*i.e.,* feature extraction) and 2) inputing the feature set into classifiers which can automatically recognize lesion patterns through machine learning techniques (*i.e.,* pattern recognition).

As an early and correct diagnosis enables a timely treatment which reduces the potential risks (*e.g.,* metastasis, disfigurement) of lesions, the possibilities of increasing the accuracy of lesion classification have been widely discussed in the computer aided skin lesion analysis field. Many works have been proposed over the past 25 years [4, 91, 10, 56, 21]. One strategy is to employ advanced screening equipment (*e.g.,* dermoscopy) as it could enhance the visualization of lesion structures and allow the observation of more lesion details. Most importantly, these systems could provide additional diagnostic features. For example, the ABCD rules [103] and 7-point Check-List [44] based features are associated with dermoscopy and these features largely improved the differential diagnosis rate of pigmented skin lesions (*i.e.,* melanoma against melanocytic nevus). The other strategy is to develop new computer vision-based techniques (*i.e.,* machine learning algorithms). In our opinion, the former is essential, as no classifier can work well with poor quality and insufficient lesion descriptors. Therefore, in order to improve the diagnosis accuracy of lesions, the top priority should be including more features carrying complementary diagnostic information.

So far, dermoscopic images are commonly used in the field. Many researches reported that this screening technique improved the diagnostic accuracy of both experienced dermatologists and computer algorithms for pigmented lesions, though it appeared to be less helpful for inexperienced clinicians [7]. The fundamental advantage of dermoscopy is that it enables the assessment of the sub-surface structure of skin through a magnification process. However, as far as the information source is considered, the dermoscopy could not produce any additional information other than the colour compared to the conventional 2D imaging systems. As a result, dermoscopy is mostly reported helpful in diagnosing pigmented lesions, for which the colour carries the most important discriminative message. However, when a broader range of lesions (*e.g.,* non-pigmented) are taken into account, extra morphological information is worth exploring.

As discussed previously (in Section 1.1), skin is the outermost tissue of the human body whose surface is characterised by polyhydric mesh structures representing the three dimensional organisation of the dermis and the subcutaneous tissue [13, 14]. Normally, this topical structure is highly regular when the skin is healthy, but becomes irregular when skin problems arise. Due to different pathogenic of different skin problems, such as the cell of origin, the outermost surface usually has different topographical appearances. For example, typical keratinocyte derived tumour Basal Cell Carcinoma (BCC) has persistent, non-healing, eroded areas with poorly defined borders. These characteristics make them very different from the unruffled, circular or oval shaped smooth landscapes of common moles. These observations suggest that the topography of the skin surface can be considered as a mirror of the functional skin status [16]. It should be deemed as another important skin descriptor in addition to colour, revealing delicate differences within lesions and playing an important role in dermatological diagnosis. As a result, we hypothesize that *the 3D shape of skin lesions embodies complementary features that serve to improve lesion identity recognition.*

To date, the relative reports on the application of 3D imaging systems to skin lesion diagnosis are very rare. There is even less reliable study of the usefulness of this addition information (*e.g.,* depth). To fill in the blank in the literature, the goal of this chapter is to give a comprehensive evaluation of our claim through the development of a multiple class lesion classification system which involves feature extraction, feature selection and pattern recognition. A stereo imaging system is applied for lesion data collection. This new sensor allows the simultaneous acquisition of 3D geometric and 2D colour information of the lesion surface. The two databases (DATABASE I and DATABASE II, detailed in Appendix D) collected from the Dermatology Department of Edinburgh University using this sensor are used in the experiment. There are five classes of lesions taken into account, including two types of skin cancers BCC (Basal Cell Carcinoma), SCC (Squamous Cell Carcinoma), as well as three kinds of benign lesions, AK (Actinic Keratosis), ML (Melanocytic nevus) and SK (Seborrheic Keratosis). The deadly form of skin cancer - melanoma, however, is not included because of the shortage of samples. None of the considered cancers are as life-threatening as melanoma. However, as they are mostly exhibited by patients in clinic, some concerns about them have caused the patient to make an inquiry. Identifying these common skin lesions optimizes the lesion selection for biopsy and pathology review and enables the correct course of action. So far, there has been almost no image analysis about these lesion conditions (BCC, SCC, AK, SK) considered in our work [56]. Besides, most works are limited to the binary discrimination problem of melanoma and benign pigmented lesions. Taking into account the inter-similarity and intra-variability between lesions, including multiple dermatologic conditions makes the correct diagnosis even more challenging [104]. From this point of view, adding extra diagnostic information appears to be more meaningful.

In the following sections, we will show that the addition of depth data increases the diagnostic accuracy for both human and computer based classifications. The performance of human diagnosis has a significant increment of 8.5% on the 3D images relative to the 2D images (Section 5.5.1). By combing depth, colour and texture features (Section 5.4.1) in a Support Machine Vector classifier (Section 5.4.3), we show that an improved classification rate of 80.67% compared to that only using colour features (75.25%) (Section 5.5.2).

5.2 Literature Review

5.2.1 Computer Based Skin Lesion Classification Systems

For the purpose of achieving an objective, consistent, quantitative and cheap diagnosis, Computer based Skin Lesion Diagnosis (CSLD) systems have been developed to mimic the presumptive diagnosis process given by dermatologists. These systems translate knowledge of dermatologists into a computer program as the means of applying medical image analysis techniques to the quantitative measurement of pathological alternations of human skin [5]. A review of the history and the development of CSLD systems can be found in Section 2.1 and Section 2.2.3.

The current progress is very encouraging. However, it must be noted that these results are obtained conditionally. First, almost all CSLD systems perform the diagnosis on a pre-selected image set. The selection is normally based on two principles: 1) only the data with acceptable image quality will be considered in the database. For example, if the images have heavy hair or the size of the lesion is too big to fit in the image or if there is insufficient contrast between the lesion and the healthy skin, the corresponding cases would be ignored [10]. As many such cases could correspond to the uncommon situations (or outlier) of certain lesion types, this pre-selection step may reduce the degree of difficulty to some extent. Though, because that the image quality heavily influences the computer aided diagnostic accuracy, this step appears to be necessary in all image based analysis systems. 2) The type of lesions considered in the CSLD systems is always very limited. In most cases, only pigmented (or melanocytic) lesions are taken into account in the lesion pool (*e.g.,* [20]). The aim of the CSLD systems is merely to distinguish melanoma (malignant) and melanocytic nevus (benign) [21]. Many non-pigmented lesions that are frequently presented in clinics are excluded and no discrimination of them is considered. At present, there are very few systems able to distinguish lesions among more than two dermatological types. It is obviously not fair to compare the computer aided diagnostic result based on such a limited database to that given by dermatologists who conduct diagnosis using a much broader range of lesions. Despite the higher performances reported in many works, at this stage, experienced dermatologists are still believed to produce a most convincing diagnostic rate, followed by the computer-based algorithms which might serve as a second opinion for inexperienced clinicians [7].

On the other hand, because the input of CSLD systems are mostly dermoscopic images, the following-up analysis has been only based on 2D colour information. In recent decades, with the rapidly growing research in 3D computer vision, the analysis of objects in 3D space becomes possible. There can be no question that the processes for depth recovery are a part of the human visual system and that they can be vital for certain tasks. However, there are arguments on whether depth information is necessary in recognition. Ultimately the choice of whether to use a 2D or 3D system is one which can only be determined by the application being researched [18]. Some early works on applying 3D systems on skin analysis are reviewed in Section 2.3.

In our work, we make improvements through 1) adding more diagnostic features, 2) performing the comparison between 2D and 3D-based features on larger databases, 3) developing a new feature selection method by integrating forward and backward search in a novel way and 4) investigating the problem using different classifiers and choosing a more suitable classifier for the task. Besides, we investigate the benefit of 3D data with human tests. In the following section, we give a detailed literature review on each related sub-topic in the classification process.

5.2.2 Feature Extraction

If one tries to get a computer to classify objects, a sound approach is to measure some prominent features of each object and to use these features as an aid to classification. From this point of view, lesion classification is a post process of feature extraction. Therefore, the extraction of good features is vital for an accurate classification [105, 47].

The feature extraction is normally based on two factors:

1) *the captured lesion data (*e.g., *images)* which relates to the computer perception and determines the kind of source (*e.g.,* colour or shape) to be analyzed. The goal of feature extraction is to quantitatively characterize the image content by computer vision approaches, *i.e.,* modifies the data from the lowest level of pixel (or voxel) data into higher-level representations.

2) *the diagnostic criteria.* Several diagnostic criteria based on dermoscopy have been proposed and tested in the clinical practice. The most commonly used criteria are the ABCD rules and the 7-point checklist (more details of these two criteria can be found in Section 2.2.2). Both of the above diagnostic criteria are dermoscopic-image based. Therefore, their derived features can only be colour or colour-texture based. Frequent research confirmed the importance of these two kinds of features in obtaining an accurate classification [27]. However, the skin lesion surface is a detailed

landscape. Its surface shape could also yield a valuable source of features on which to base classification. Therefore, it is important to assess the vertical growth related features, as well as the roughness and indented aspect. The limitation of the imaging system makes it impossible to assess this potential diagnostic information. As a result, very few skin lesion diagnostic applications have considered using the surface shape properties derived from 3D data, although the authors in [15] noticed the importance of surface shape variables. They characterized the shape features by irregularity variables. However, as these properties are extracted on the graphic three-dimensional pseudoelevation anaglyph developed from the colour image rather than the real depth image, they were still colour texture instead of topographical structure based features. A single intensity image proves of limited use, as pixel values are related to surface geometry only indirectly, that is through the optical and geometrical properties of the surfaces as well as the illumination conditions. Hence, it is preferred to acquire images encoding shape directly. Castellini *et al.* [53] is possibly the first group carrying out the real 3D measurement of the superficial structure of the skin lesion. As the second growth phase of the melanoma, vertical growth is considered as an important clinical prognostic information, the authors were inspired to access the lesion height measured using a laser triangulation technique, which had the disadvantage of long capturing time. Even though their 3D measurement system enhanced the knowledge in the field of measurement and reconstruction of skin characteristics, their work mainly focused on proving the measuring ability of the system to capture the morphological characteristics of the lesion. There was no further discussion about the diagnostic value of the data. The idea of including surface shape based properties into the lesion diagnosis has been dropped behind because of lack of good 3D imaging systems until recent years. In [47], the diagnosis criterion was based on the assumption that the melanoma surface had more irregularity in 3D shape than benign lesions. The photometric stereo allows the capturing of the shape of lesion in 3D format and allowed the author to investigate the effectiveness of these 3D-based texture features (in terms of 4 curvature pattern based properties) in melanoma diagnosis. A test on a small-scale data set comprised of 23 melanoma and 53 benign lesions indicated the effectiveness of the 3D curvature pattern in melanoma diagnosis, though the improvement was without sufficient statistical proof when compared to the classic 2D features. The author also pointed out that only using 3D shape the results were not completely reliable and other indicators should be taken into consideration as well. McDonagh *et al.* [56] made preliminary investigations into the simultaneous use of colour plus 3D based-properties for lesion diagnosis.

Several features were designed to account for the 3D related features, such as height, roughness, *etc.* Their experiment showed the 3D based features would be selected in the greedy forward feature selection process and the classification result suggested that incorporating topographical features provided better diagnostic results than those obtained merely using colour texture features, though also without sufficient statistical support. For more information about the features utilized for skin lesion characterization, we refer the reader to [2].

5.2.3 Feature Selection

The task of a classifier is to use features to assign the represented object to a category or class [106]. However, the diagnostic accuracy is not necessarily improved with the increasing number of the features. With the number of features increasing, the classification rate of a classifier normally decreases after a peak [2]. A high dimensional feature set as the input of classifiers may turn them inefficient and even make them inapplicable, such as the singularity problem in Quadratic Discriminant Analysis (QDA) when the dimension of features exceeds the number of samples. Besides, redundant and irrelevant data decreases the classification ability and leads to false conclusions [49]. In [106], the authors proved the usefulness of feature selection on improving the classifier robustness and performance through a rigorous empirical study. They addressed that feature selection helped to focus the attention of a classification algorithm on those features that were most relevant to predict the class and improved the accuracy, efficiency, applicability and understandability of a learning process and its resulting model.

Generally, feature selection algorithms can be classified into two main categories according to their evaluation criteria: filters and wrappers. Filter approaches rely on general characteristics of the data to select a subset of features without involving any learning algorithm. Instead, they ‘filter’ out irrelevant and redundant noisy features. The Principal Component Analysis (PCA) based feature selection belongs to this category and it is commonly used in feature set reduction because it can reveal composite features that are more effective than their individual constituents [107]. For example, Lyatomi *et al.* [108] used PCA to reduce a total of 428 image features into 198 orthogonal principal components by sub-space projection. However, PCA does not allow one to observe the relationship between features and the patterns, such as which kind of features are informative for representing a pattern.

Wrapper methods are computationally expensive but are better suited to classification tasks [10]. They use the prediction performance of a pre-determined learning algorithm to evaluate the goodness of feature subsets [109]. Wrapper feature selection can be characterized through the search strategy employed. A direct solution would be an exhaustive search for all possible combinations. However, the main drawback is that it has complexity of $O(2^D)$ for D features [110]. Therefore, finding the optimal combination generally costs a lot computing time. For example, trying all these feature combinations in a 48 dimensional feature space would result in $\sum_{D=1}^{48} choose(48, D) = 2^{48} - 1$ times searches, where D is the size of feature subset. Given 15 seconds per combination, this is computationally infeasible. A naive simplified approach is to rank the features and choose the top D to create the best subset. But this procedure overlooks the possibility of 1) features with poor individual contributions performing better in combination because of carrying complementary information [106] and 2) top features performing poorly in tandem because of redundancy. A widely acceptable alternative is the greedy search strategy, including sequential forward and backward strategies. The Sequential Forward Selection (SFS) algorithm is efficient but suboptimal. It starts with an empty feature set and iteratively adds a single feature that could optimize the classifier performance when combined with all those previously selected features. SFS continues until a required dimensionality is achieved or an evaluation criterion is reached. In [56], the feature space was reduced from 30 to 10 using this strategy. It rejects statistically negligible features during incremental selection, so that those highly correlated features are automatically excluded from the feature set. This approach is very sensitive to the first chosen feature. The Sequential Backward Selection (SBS) is similar to SFS but works in the opposite direction. It initializes with the full feature set. At each step, SBS finds a single feature, by removing which the classifier performance improves the most significantly [110]. Both of these two methods solve the redundancy problem in the naive ranking approach and have much lower computational burden compared to the exhaustive search approach, especially the SFS. However, the common problem with these two approaches is they often end up with sub-optimal results because of the inability to re-evaluate the usefulness of features that were previously added or discarded. A solution is the Sequential Floating Forward Selection (SFFS), which dynamically integrates the forward and backward selection in control of certain evaluation metric. The common SFFS starts with a forward selection. Once the improvement is less than a pre-set threshold, the backward selection is switched on as long as a better subset than those of the same size obtained so far is found, or vice versa.

5.2.4 Classifiers

The task of a supervised classifier is to use feature vectors to assign the observed object to a category or class [106]. This is achieved by producing a learning model from a labeled training set. Various successful techniques have been proposed to solve the problem. Some classical methods are K-Nearest Neighbour (K-NN), Artificial Neural Networks (ANN), Discriminate Analysis (QDA or LDA) and Support Vector Machine (SVM).

K-NN is one of the oldest non-parametric classification algorithms [111]. The test data (or unknown feature vector) is assigned to the class that occurs most often in the set of K-Neighbours. The problem of K-NN is that large numbers of training set patterns are normally required for achieving a low error rate. This leads to significant storage and computation problems.

A more generalized approach called Bayesian Classifiers is based upon the principle of Maximum A Posteriori (MAP). This statistical based method is the most classical recognition paradigm used in skin lesion diagnosis and its major problem is the need for large learning samples [2]. But the advantage of this classifier is that it is straightforward and does not introduce any parameters to be tuned. Given a problem with K classes $\{C_1,\ldots,C_k\}$, the label of an unknown sample with feature $X=(x_1,\ldots,x_N)$ should be assigned to class i, if $p(C_i|X) > p(C_j|X)$, for all $j \neq i$. Using the Bayes's theorem, this could be further simplified as $p(X|C_i)p(C_i) > p(X|C_j)p(C_j)$. By extending the probability function (using a Gaussian distribution model), the above inequality generates the discriminant function or decision rule. One can see that, to find the class of an object one needs to know two sets of information: 1) the basic probability that a particular class might arise (as known as the priori probability $p(C_i)$) and 2) the distribution of values of features for each class (also known as the class-conditional density $p(X|C_i)$). Each information can be found straightforwardly by observing the training set. Prior probabilities reveal the frequency of individual cases in the real world and it can be estimated directly from the training set as the fraction of the training set data points in each class. However, sometimes, the expected prior probabilities differ from those represented by the training set (*e.g.,* the dataset where samples in each class are equally distributed artifically). This often happens in the medical research field where the proportion of disease cases is normally small and needs to be artificially increased in order to obtain a good variety for further analysis. In this context, the prior probabilities in the training and testing datasets differ from each other. A simple solution to compensate for the different priors suggested by Bishop [112] is to replace the prior probabilities estimated from the training dataset by the ones obtained from medical statistics in the general population. The distribution of feature values is often modeled using a multivariate Gaussian distribution, for which two parameters, the mean and the covariance need to be estimated. Different parameter estimations result in different discriminant function (*e.g.,* QDA, LDA).

The Support Vector Machines (SVM) is a more advanced method. It is based upon the idea of maximizing the margin, *i.e.,* maximizing the minimum distance from the separating hyperplane to the nearest example [111]. To date, SVM is among the most robust and successful classification algorithms in the field. However, the disadvantage of SVM is that there are several choices to make, because the effectiveness of SVM depends on the selection of kernel, the kernel's parameters and some other parameters(*e.g.,* soft margin).

Both K-NN and Bayesian Classifiers can be directly extended to multiple label cases. The basic SVM supports only binary classification, but it can be extended to multiple classifications by reducing the multi-class problem into a set of binary classification problems through different formulations. A good review can be found in [113].

For the skin lesion classification, various classifiers have been applied. Dreiseitl *et al.* [114] conducted a comparison of the discriminatory power of the six main classifier categories on the task of classifying skin lesions. Their result showed that logistic regression, ANN, and SVM performed on about the same level, with K-nearest neighbours and decision trees performing worse. However, there has no standard method in the application of skin lesion diagnosis up to date. Ganster *et al.* [41] performed a three category (benign, dysplastic and malignant) lesion diagnosis using a K-NN classifier with K assigned to 24. McDonagh *et al.* [56] used a Bayes Classifier with a unimodal multidimensional Gaussian model for the multiple lesion classification task. In [28], the authors put the extracted features into a back-propagation neural network and achieved 95% diagnostic accuracy in the automatic discrimination between melanoma and benign lesions (nevus). German *et al.* [32] performed the learning and classification stage using AdaBoost.M1 with C4.5 decision trees which gave promising classification results that were superior than those reported in the literature.

5.3 Prior Study of the Influence of 3D Data on Human Diagnosis

When asking dermatologists about the usage of the surface shape in lesion diagnosis, they could not provide a decisive answer [1], although their professional intuition makes them feel that the shape might be helpful. So far, there has been no related experimental study on this subject. In this section, we propose an experiment to answer this question as to whether there is any 'benefit' of the 3D data on humans' diagnosis. The overall Diagnostic Accuracy (DA) is employed as an evaluation metric, which can be later used as a baseline to compare with that produced by computer algorithms.

Objects

The experimental objects are 100 lesion images, which have the same size and resolution, as well as adequate 3D models. They were evenly selected from five lesion classes: BCC, SCC, AK, ML and SK. In each experiment, 40 images, two batches of 20 images would be presented to the evaluator. The 20 images in the first half of each experiment were always different from the second 20 images. Each batch of 20 images was selected by stratified random sampling so there were four images from each of the five classes. Users could have interactions with both 2D and 3D images in terms of rotation and zooming.

Subjects

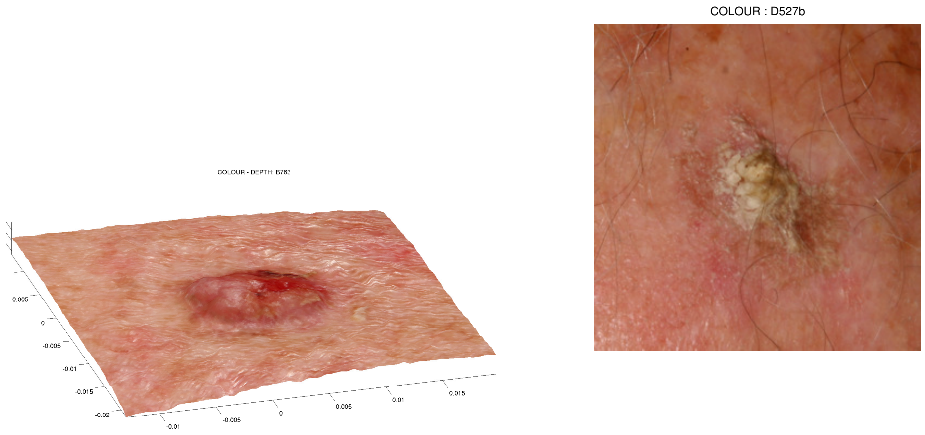
The experimental subjects are medical students who were undertaking dermatology course training. But they had not been exposed to any 3D images or attended 3D-image based diagnosis training before the experiment. These students were organized in 6 groups and attended the experiment at different time slots. They were also asked to perform different operations.

1. **Group 1**

   Group 1 is comprised of 50 student volunteers who attended the experiment 1 week before Sept 2010 exam. Each student attempted to diagnose 20 2D images and then 20 different 3D images. An example of a pair of 2D and 3D images is shown in *Fig.* 5.1. In this experiment, the question is *whether there is an improvement on DA by using 3D images*.

2. **Group 2**

[1] Private discussion with Dr. Jonathan L Rees, the Grant Chair of Dermatology at the University of Edinburgh



(a) 3D image

(b) 2D image

Figure 5.1: 2D and 3D images used in human performance test

Group 2 is comprised of 13 students who attended the assessment on the final day of 2-week attachment in Nov 2010. They were asked to repeat the Group 1 experiment as the question is raised as *whether the findings in experiment 1 will hold true on a second group*.

3. **Group 3**

   Group 3 is comprised of 13 students who also attended the assessment on the final day of 2-week attachment in Nov 2010. Each student attempted to diagnose 20 3D images first and then 20 different 2D images. This is in fact a reverse experiment of Group 1&2 experiments. From the result, one can answer *whether the previous findings hold true if they see the 3D images first (or whether the order of exposure to 2D and 3D images carries an effect on findings)*.

4. **Group 4**

   This group includes 14 students who attended the assessment on final day of 2-week attachment Dec 2010. They were asked to diagnose 20 2D images then 20 different 2D images as it is worth knowing that *do the students improve on the second 20 images irrespective of 3D*.

5. **Group 5**

   Group 5 is comprised of 14 students who attended the assessment on the final day of 2-week attachment in Jan 2011. Each student diagnosed 20 2D fixed images and then 2D images with ability to rotate. This experiment intends to answer *whether the interaction (here means rotation) of the 3D images but rather than the 3D images themselves improves their DA.*

6. **Group 6**
   The last group is comprised of 15 students who also attended the assessment on the final day of 2-week attachment in Jan 2011. Each student diagnosed 20 2D fixed images and then 2D images with ability to zoom. This experiment intends to answer that *does the interaction of the 3D images with respect to scaling improves the DA.*

The diagnostic results of each group were recorded and marked by Dr. Ben Aldridge, from the Dermatology Department of Edinburgh University.

5.4 Methods

To evaluate whether the addition of 3D depth information would potentially benefit in the diagnosis rate relative to only using colour information, one needs to develop an automated pattern recognition system. The three main techniques used for this pattern recognition process are: 1) extracting morphological features that are indicative of skin conditions and allow classifying different conditions by type from colour and depth data, 2) selecting the optimal combination of features from the extracted feature set and 3) choosing a suitable classification model for lesion recognition.

5.4.1 Lesion Descriptors

The lesion descriptors are extracted based on a preliminary study in our lab, in which a total of 30 features were extracted from the colour (22 features) and depth (8 features) image data. More details can be found in [115]. In this work, the feature set is further extended to a set of 48 comprising of 34 colour and 14 depth properties (see *Table.* 5.7 for the list of features), which can be summarized into two categories:

1. **Global Properties**
   In the Segmentation Chapter 4, each lesion is decomposed into two sub-regions: the interior (lesion region) and the outer (healthy skin region) of the boundary. Each region could be characterized by some statistics of its distribution (*e.g.,* mean and variance per channel in RGB and depth). Hence, a family of first-order statistics based features called *Relative Colour Brightness Features*

(9) are proposed to calculates the ratios of colour intensities within the lesion relative to the normal skin. They are formed as $\frac{\mu_{a,S}}{\mu_{b,L}}$, where μ is the mean value of colour (*e.g.,* R,G,B) channel and $\{\mathrm{S, L}\}$ denote the $\{\text{skin and lesion}\}$ patches, respectively. The *Relative Variability Features* features (4) are extracted based on the second-order statistics and have the form of $\frac{\sigma_{c,S}}{\sigma_{c,L}}$. This group of properties assesses the variability of colours inside the lesion relative to the normal skin patch and attempts to automatically compute quantities which emulate the 'colour' aspect of the ABCD clinical diagnosis rule discussed previously. It could be directly applied onto the depth image as $\frac{\sigma_{d,S}}{\sigma_{d,L}}$. Furthermore, the data (intensity or depth) distribution of a lesion can be measured using the skewness (the absolute value) and kurtosis properties based on the histogram of each colour channel or depth image. This family is comprised of *Absolute Skewness(Kurtosis)* ($|Skewness(Kurtosis)|_{c,L}$) (8), *Relative Skewness(Kurtosis)* ($\frac{|Skewness(Kurtosis)|_{c,L}}{|Skewness(Kurtosis)|_{c,S}}$) (8). These types of feature measure the symmetry and the flatness of the data distribution and helps to spot irregular surface shape. Another three global features are the lesion size descriptors, which are also considered as diagnostic criteria in the ABCD rule. They are area, average height of lesion and volume of a lesion. The area is calculated based on the detected contour. The volume is calculated based on the boundary and the height information derived from the 3D data. As most lesions have a very irregular shape, the average height of a lesion is only an approximation of the ratio of the volume and the area. In addition, to characterise the mass distribution of the lesion volume, three *3D Shape Moment Invariant Features* (3) are also implemented.

2. **Local Properties**

   Some localized features of texture and colour distribution were also extracted as *Peak and Pit Density Features* (12). Image data is first convolved with a Gaussian filter with a certain scale (σ) to remove fine textures. Based on the smoothing result, a local peak/pit is defined as a pixel whose value was larger/smaller than the eight nearest neighbours. The ratio $\frac{\#peaks_{c,\sigma}+\#pits_{c,\sigma}}{Area}$ is computed to account for the local texture distribution, where $c \in R, G, B, d$ and $\sigma \in 0.5, 1.0, 2.0$ (denotes the Gaussian filter standard deviation).

Many more features exist in the literature, *e.g.,* the co-occurance features in [50]. However, our work is not on the issue of exploring as much as features as possible but rather to investigate the importance of depth related features. Therefore, instead of considering more features, we keep using most of McDongh's features [56] (see *Table.* 5.7) and adding some histogram based features. These features form a feature vector which represents the characteristics of a particular lesion. To avoid the bias caused by the varying range on different components of the feature vector, a range normalization operation is performed to transform each feature component to be zero-mean and unit variance over the whole data set.

5.4.2 Feature Selection

The feature vector extracted in the previous section cannot be directly fed to the classification with respect to factors like redundancy. To reduce the feature dimension and find the optimal feature subset, we propose a novel selection strategy that integrated forward and backward selection. As mentioned in Section 5.2.3, in each backward step, the Sequential Floating Forward Selection (SFFS) strategy finds and deletes one feature f^-, without which the rest of the features can produce better performance when compared with deleting other features. However, in our application, we found this method could easily fall into a repeating loop. For example, a feature removed in the backward selection step could be re-selected in the forward selection step and thereby the selection process does not move forward. To solve this problem, we propose the Sequential Pair-wise Feature Selection (SPFS) which can enlarge the search space compared to SFFS. In each backward step, instead of only considering one feature, a pair of features f^+ and f^- has to be found. The replacement of f^+ with f^- should improve the classifier performance compared to other combinations with the same feature subset size. The Pseudocode for this approach can be found below:

Pseudocode of Sequential Pair-wise Feature Selection (SPFS)

```
While $k \leq K$ % $K$ is a pre-set feature number
1     Initialize feature set $F_0 = \{\emptyset\}$, $k = 0$
% Forward feature selection step
2     Select the $k^{th}$ feature $f^+$ by
         $f^+ = \arg\max_{x \notin F_k} [DA(F_k + x)]$
3     Update $F_k$
         $F_k = F_k \cup \{f^+\}$; $k = k + 1$
4     If $DA(F_k) - DA(F_{k-1}) > \varepsilon$ (e.g., $\varepsilon = 0$)
         Go to Step 2
      Else
% Backward feature selection step
5        Find the best pair of features $f^+$, $f^-$ (only
         one round) by
         $\{f^+, f^-\} = \arg\max_{x \in F_k, y \notin F_k} [DA(F_k - x + y)]$
6        Update $F_{k'} = F_k \backslash \{f^-\} \cup \{f^+\}$
7        If $DA(F_{k'}) > DA(F_k)$
      % If backward step results in no improvement
      % keep the forward selection result; otherwise
      % replace $F_k$ by $F_{k'}$
            $F_k = F_{k'}$
8        Go to Step 2 % forward selection step
END_WHILE
```

It is worth noting that in each backward selection step, the feature f^- belongs to the selected feature set F_k which is determined in the latest forward selection step. The replacement f^+ is chosen from the feature set that excludes the features previously selected (*i.e.*, F_k).

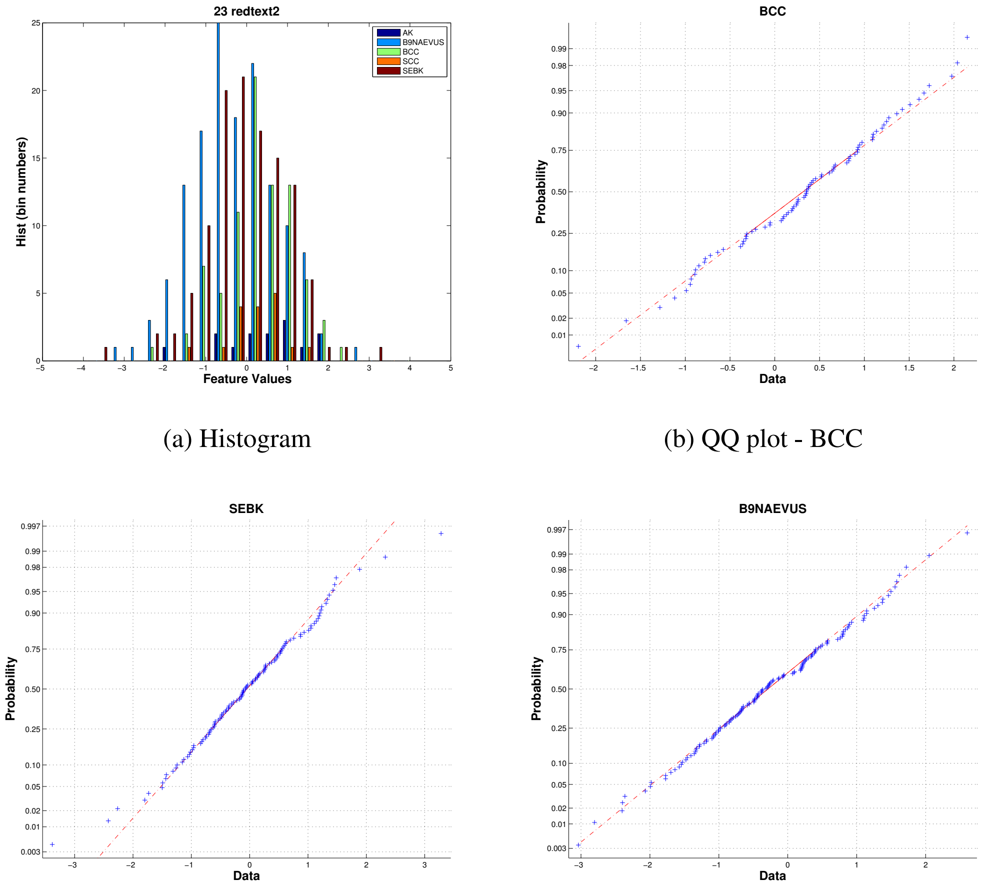
The feature evaluation criterion is the lesion Diagnostic Accuracy (DA). The selection process terminates when a pre-set number of feature size is reached. The optimal size of feature subset space (D) is the turning point where the performance of classifier starts to declines with the addition of new features. At this stage, a D dimensional feature vector is created. This sub-selected and normalized feature vector can thereby be fed to a classifier, which ensembles to obtain a classification model that can differentiate different types of lesion.

5.4.3 Classifiers

Selecting a suitable classifier has large impact on the final performance of lesion diagnosis. For a given classification task, a final selection of the 'best model' should be based on the empirical comparison of the classification performances of different classifiers [114]. Therefore, we implement three standard classification methods, which are K-NN, Bayes classifier and SVM.

Designing a Bayes classifier entails both the process of choosing the form for the probability density functions for each class and the process of choosing the parameters which describe the density function. According to visual examination, most feature value distributions can be modeled by the unimodal Gaussian density function (see *Fig.* 5.2). Thereby, the unimodal multivariate Gaussian observation model is chosen for the Bayes classifier.

The Gaussian distribution model parameters - the mean and covariance are estimated using four different ways from the training data. This results in four sub-models: LDA, QDA, DQDA (as known as Naive Bayesian) and DLDA. QDA is the short for quadratic discriminate function, in which the covariance of each class is estimated from the samples in the corresponding class. Normally, these covariance matrices are different (*i.e.,* $\Sigma_i \neq \Sigma_j$, for $i \neq j$). However, once the dimension of feature set is larger than the number of sample in a certain class, the estimation of the Gaussian parameters (covariance) can be ill-posed. This limits the size of feature subset. To avoid this problem, one can assume that the class covariances are identical (*i.e.,* $\Sigma_i = \Sigma_j = \Sigma$). This is the so-called Linear Discriminant Analysis (LDA). The dimension of the feature subset space can be as large as the database size. When the covariance matrix is further simplified as a diagonal covariance matrix $\triangle = diag(\sigma_1^2, \sigma_2^2, \ldots, \sigma_D^2)$, one can obtain the Diagonal Linear Discriminant Function (DLDA). The conceptually simple approach of LDA and its sibling, DLDA (where all classes use the same diagonal variance matrix), remain among the most effective procedures in the domain of high-dimensional prediction [116]. Another solution to solve the ill-posed problem is DQDA, which is sometimes called the 'naive Bayes Classifier'. The 'naive Bayes' assumes independent covariance and it often works well in small sample and high feature space



(a) Histogram

(b) QQ plot - BCC

(c) QQ plot - SK

(d) QQ plot - ML

Figure 5.2: Features in five lesion groups (*e.g.,* the Relative variability feature of red colour channel) roughly obey the Gaussian distribution as shown in *Fig.* 5.2(a). The qq-plots of feature 23 (*Fig.* 5.2(b), (c), (d)) also indicate that the data follow the Gaussian distribution as plots in the different lesion groups are close to linear, though with outliers evident at the high/low end of the range.

situations. For these Bayes models, to avoid the bias caused by the unbalanced sample distribution, the *a priori* class probability for each class should be estimated using the incidence rates in the training data.

We implement all these four models for the Bayes classifier. Hence, in total, six classifiers are taken into account in our comparison. The discriminatory performances of all these classification models are compared through classifying five types of skin lesion (*Table.* D.1 shows the number of lesions for each class). The optimal feature subset for each classifier is automatically determined using our Sequential Pair-wise

Feature Selection (SPFS) method, respectively. Because of the relatively small data set of certain lesions (*e.g.,* AK and SCC), the database is not separated into independent training and test sets. Instead, a leave-one-out cross-validation is used for training and testing set generation. This approach trains each classifier on all of the available skin lesions apart from the one that is to be classified. Take a database including N samples for instance, the system is trained N times on all data except for one sample and a prediction is made for that sample. This affords us the maximum mileage possible from the available data in terms of model training [56].

5.5 Experiments and Results

5.5.1 Whether the 3D Data Improves Humans' Diagnosis?

Group	Image Batch	Min Score	Max Score	Avg Score	Sum Scores	Total Test Images	DA%	Paired wilcoxon test (p value)
1: 3D benefit	2D	1	13	8	385	1000	38.5	2.148e-06 (significant)
	3D	5	14	9.5	470	1000	47.0	
2: Repeat	2D	3	10	5	77	260	29.6	0.01004 (significant)
	3D	4	12	8	103	260	39.6	
3: Reverse	3D	5	12	8	109	260	41.2	0.01025 (significant)
	2D	4	10	6	85	260	32.7	
4: Order effect	2D	1	10	6.5	82	280	29.3	0.7523 (not significant)
	2D	2	9	6	84	280	30.0	
5: Rot effect	2D	1	10	5	78	280	27.9	0.3905 (not significant)
	2D Rot	2	9	6.5	84	280	30.0	
6: Zoom effect	2D	2	10	7	99	300	33.0	0.5263 (not significant)
	2D Zoom	4	10	6	104	300	34.7	

Table 5.1: Human Diagnosis Results. This table lists the results of the experiments on the six groups. For each group, results for both batches are recorded, including the Min Score, Max Score, Avg Score, Sum Scores and Diagnostic Accuracy (DA). The p value of the paired wilcoxon test between two batches is also given. It measures the statistical difference between the diagnosis over two batches

The diagnosis results from Group 1 to Group 6 are listed in *Table.* 5.1. Group 1 demonstrates a primary test, which compares the Diagnostic Accuracy (DA) between using 3D and 2D images. The DA distribution of the two batches are demonstrated in *Fig.* 5.3. It shows that the students who diagnose on 3D images achieve an average DA of 47.0%, which is much improved compared to that based on 2D images, which is 38.5%. The paired wilcoxon test [117] suggests that this improvement is statistically significant with respect to a p value of 2.148e-06. One can also see that the diagnostic variation between each batch is smaller for 3D than 2D. The minimal diagnostic score is 1 out of 20 for 2D images and 5 out of 20 for 3D images; while the maximal diagnostic score is 13 out of 20 for 2D images and 14 out of 20 for 3D images. These might indicate that the 3D images based diagnosis appears to be more reliable and robust. The experiment Group 2 is a repeat test of Group 1 with different images. From *Table.* 5.1, the same conclusion holds, though the average DAs of both 2D and 3D images are smaller compared to the respective values in Group 1. This phenomena exists in all following groups. It could be explained by the time factor, as Group 1 test is only taken a week before exam while the others are at least 2 months after the exam. This could be also because that only the students in Group 1 volunteered.

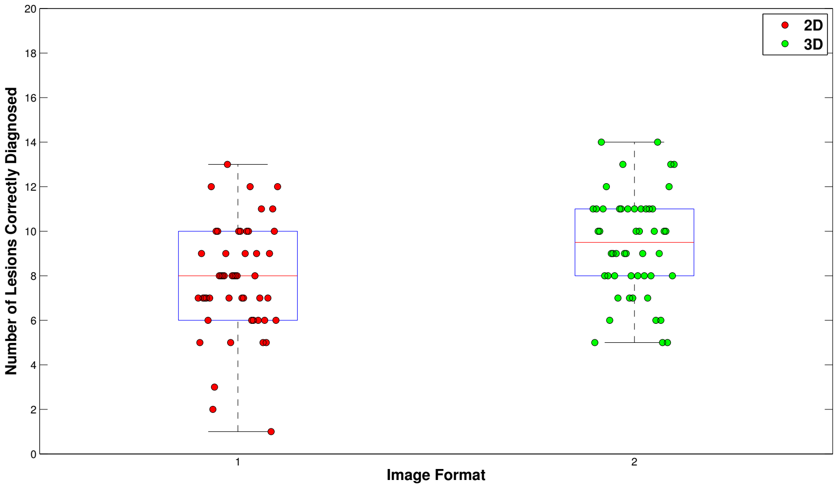


Figure 5.3: Human diagnosis based on 2D and 3D images. The red and green spots indicate the number of lesions being correctly diagnosed by individual raters. The red ones relate to the diagnosis on 2D images and the green ones relate to that of 3D images

The question of whether the significant improvement of 3D over 2D is related to other side effects is tackled by experiments on Groups 3, 4, 5, 6. The reverse test of Group 3 shows that the order of 2D and 3D images does not influence the conclusion. Group 4 experiment further indicates that there is no order effect, because there is no difference between the diagnosis based on two 2D image batches displayed in succession. The last two tests answer the question of whether the interactions with images would affect the DA. According to the result, the diagnosis with and without the abilities of rotation and scaling have very close DAs. Therefore, all these possible factors can be excluded.

In summary, there is significantly higher DA in 3D images for Group 1, Group 2 and Group 3. This appears to be a genuine 3D effect rather than interaction or order related effect. In addition, a questionnaire-based enquiry reveals that 62 out of 76 (81%) students who attended the test found it was easier to diagnose on 3D images and 73 out of 76 (96%) preferred using 3D for teaching as they found 3D provides more information than 2D images. As a result, we have a reason to believe that there is truly a benefit of 3D images over 2D images on humans' diagnosis. In this context, it is worth considering to use the extra information carried by 3D images on computer-based diagnosis.

5.5.2 Classification Results

The aim of this section is to gain some insights into the contribution of 3D-based features towards lesion diagnosis. Our evaluation is based on Database I, which includes 369 samples over the five classes (including AK (14), BCC (140), SCC (17), ML (83), SK (115)). More details of Database I can be found in Appendix D. Three classifiers are implemented and used as the comparison platforms. They are K-NN, Bayes and SVM. In the K-NN classifier, the number of nearest neighbours K is set to two based on empirical tests. The distance is the Euclidean distance. For the Bayes classifier, four probability density parameter (covariance matrix) estimation methods produce four models, which are QDA, LDA, DQDA, DLDA. The prior probabilities for individual classes are directly estimated from the training dataset which reflect the frequencies of different lesion incidences in the hospital. For the SVM classifier, we use the libSVM package [118], where the RBF kernel is chosen and the corresponding parameters are estimated using its own functions. For training and testing set generation, a leave-one-out cross-validation is used because the shortage of samples in some classes like AK and SCC. For each classifier, its optimal feature subset is identified using our proposed Sequential Pair-wise Feature Selection (SPFS) approach. As a result, the optimal feature subset and the dimension of this subset for each classification model might vary. But in this way, the comparison is based on the best performance of each classifier. This is for the purpose of a fair comparison.

The classifier evaluation metric is the Diagnostic Accuracy (DA). It is commonly calculated as the Overall Classification Rate (OCR, *i.e.,* dividing the total correct classifications by total classifications). This is the main evaluation criterion that we use in our work (in the classifier comparison and the feature selection step). Although there are concerns that the OCR may bias the results towards the classification rates of the larger classes when using an unbalanced database. Some works prefer the Average Classification Rate (ACR) for each lesion class (*i.e.,* the classification rate is computed for each of the five classes and the average classification rate is the average of these 5 rates). However, as we need to compare the computer based classification to the human diagnosis, whose marking is based on the OCR, so that it is more fair to compare two results on the same standard as OCR. We also calculated the ACR for the classifiers comparison.

The experiment results are listed below:

1. **Comparison between 2D, 3D and 2D+3D Features**

   Table. 5.2 and *Table.* 5.3 show the classification results of each classifier using different feature combinations (*i.e.,* colour, depth and integrated colour and depth). For the convenience of future reference, let s_1, s_2 and s_3 be the classification using Colour features only, Depth and Colour features and Depth feature only, respectively. Based on the diagnostic accuracy of classifiers, combining 2D (colour) and 3D (depth) based features outperforms the others. This conclusion is held for all six classifiers.

   For each classifier, the diagnostic accuracy metrics OCR and ACR are calculated on the final result of the leave-one-out cross valuation based on the optimal feature subset selected using SPFS, which is shown in *Table.* 5.3 (the 3D-based properties are in **bold**). Different classifiers normally have different feature subsets. This might be because different classifier models fit in different feature combinations. Also, even though integrated with the backward selection, the SPFS is still sensitive to the first selected feature, which gives a large influence on the features selected afterward.

Feature Set	**Diagnostic Accuracy of Classifiers - OCR (ACR) %**							
	DATABASE I						**DATABASE II**	
	K-NN	**Bayes**				SVM		
		QDA	LDA	DQDA	DLDA	FS I		FS II
s_3: 3D	57.18(44.98)	58.27(40.55)	57.99(44.71)	53.39(38.58)	54.74(42.74)	57.18(48.14)	N/A	
s_1: 2D	71.27(55.71)	79.13(60.44)	79.40(62.20)	68.02(57.69)	67.75(51.14)	76.69(62.92)	73.28(61.30)	75.37(61.89)
s_2: 2D+3D	72.36(63.73)	79.95 (59.05)	81.57(69.51)	69.92(60.64)	69.65(63.62)	82.38(71.97)	77.22(65.41)	80.54(68.59)

Table 5.2: Diagnostic accuracy using different feature sets. FS I and FS II denote the feature sets selected using SVM classifier based on DATABASE I and DATABASE II, respectively

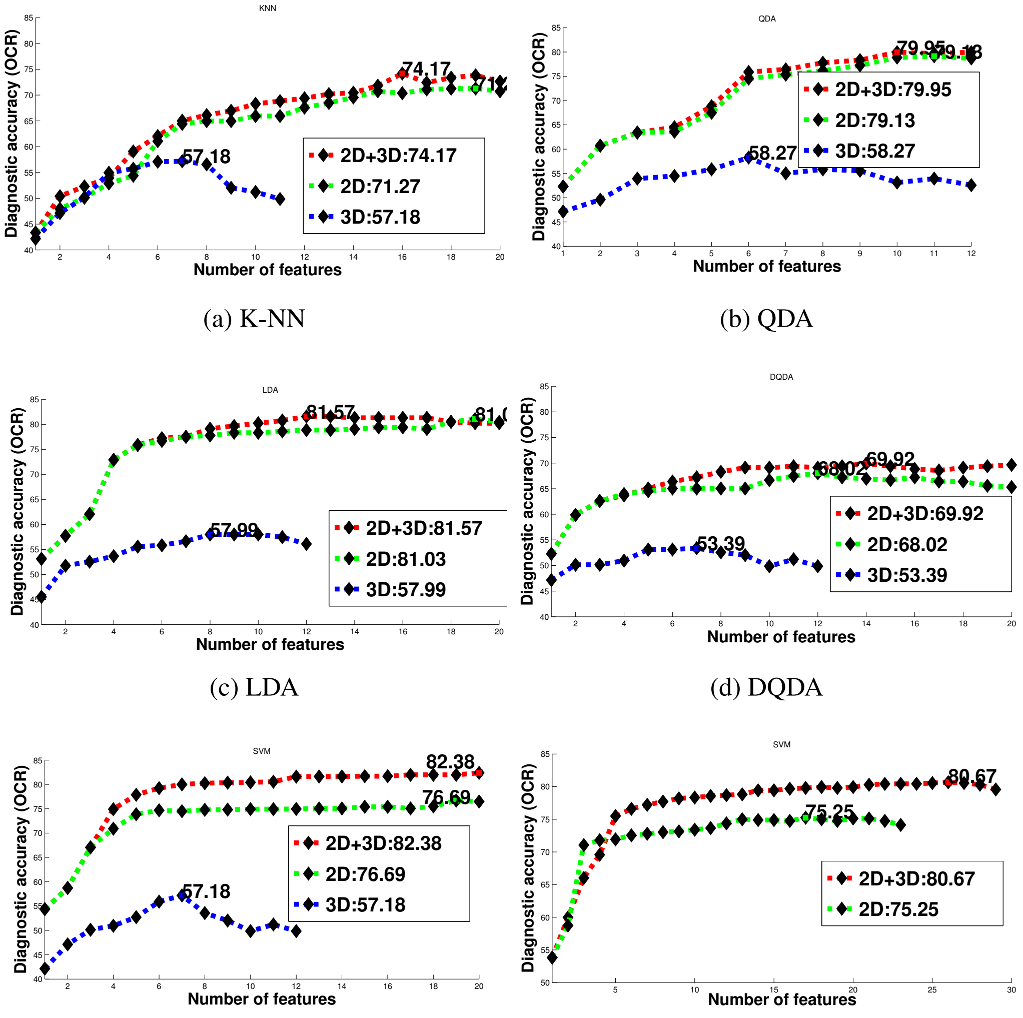
   The relationship curves between the OCR and the size of feature subset D are given in *Fig.* 5.4. For each classifier, three curves are shown. The red one rep-

K-NN	Feature set	Feature pool	Feature subset
	I	Colour only	25 11 4 22 23 18 1 31 24 9 2 12 2 6 5 7 6 3 8
	II	Colour and depth	25 **41** 27 11 8 21 1 9 5 23 31 **40** 7 **46** **43** 29
QDA	Feature set	Feature pool	Feature subset
	I	Colour only	33 21 5 9 1 19 24 28 16 11 17 20
	II	Colour and depth	33 21 5 9 1 **44** 16 **35** 13 22
LDA	Feature set	Feature pool	Feature subset
	I	Colour only	6 5 1 8 22 23 12 11 14 31 3 13 19 7 30
	II	Colour and depth	6 5 1 8 **36** 22 10 21 **39** **40** 31 23 33
DQDA	Feature set	Feature pool	Feature subset
	I	Colour only	21 12 4 29 9 2 22 1 17 25 31 6
	II	Colour and depth	33 21 12 **36** 9 **44** **45** 18 6 4 2 **35** 3 **39**
DLDA	Feature set	Feature pool	Feature subset
	I	Colour only	6 11 1 22 25 14 30 31 34 23 10 32 26
	II	Colour and depth	6 11 17 **36** **45** 21 **43** 33 23 42 34 2 22 9 **40** 10 18
SVM (DB I)	Feature set	Feature pool	Feature subset
	I	Colour only	9 5 33 1 19 10 6 31 22 32 28 16 11 7 3 24 2 30
	II	Colour and depth	9 5 33 **41** 20 **33** 28 23 **47** 11 15 18 **45** **37** 2 3 12 **38** **43**
SVM (DB II)	Feature set	Feature pool	Feature subset
	I	Colour only	9 1 5 8 4 20 21 6 35 22 31 28 16 19 7 3 24 2
	II	Colour and depth	9 **38** 21 5 1 8 23 **44** **47** 6 **45** **37** 2 3 15 20 18 12 **41** **43** 19 **42** 22 27 35 14

Table 5.3: Selected feature subset for each classifier. DB I and DB II denote DATABASE I and DATABASE II, respectively. The features derived from 3D data are highlighted in ***bold***.

resents the OCRs of s_2. It is normally above the green curve that indicates the result of s_1. Both of these two are significantly better than that of s_3 (shown in blue). However, the depth features also produce an accuracy above 50% with OCR and 40% with ACR. These findings indicate that 1) colour is the primary cue for lesion discrimination, 2) depth is also informative for lesion diagnosis, although it is not reliable when used alone and 3) adding the depth with colour increases the discrimination abilities of all classifiers. These conclusions are derived based on experiments with DATABASE I.

To confirm the findings, we repeat the feature set comparison experiment on DATABASE II (see Appendix D). The SVM classifier that performed best on DATABASE I experiments is employed. The results can be found in the last two columns of *Table.* 5.2 and *Fig.* 5.4(f). First, we directly use the feature set selected by SVM classifier on DATABASE I. The diagnostic accuracy is improved from 73.28% to 77.22% for OCR and from 61.30% to 65.41% for ACR because of the additional 3D based features. Second, in order to obtain the best result, we re-select the feature set based on DATABASE II. The new feature set is shown in *Table.* 5.3. There are eight 3D-based features selected by the



(a) K-NN

(b) QDA

(c) LDA

(d) DQDA

(e) SVM (FS I + DATABASE I)

(f) SVM (FS I+DATABASE II)

Figure 5.4: Diagnostic property comparison using different classifiers

SPFS. The confusion matrix for different feature sets (s_1 and s_2) are shown in *Table.* 5.4 and *Table.* 5.5. Again, the diagnostic rate is increased by around 5% (from 75.37% to 80.54%) for OCR and 7% (from 61.89% to 68.59%) for ACR, with the additional 3D features.

Most melanoma classification results reported in the literature reported their performances in terms of specificity and sensitivity. A survey of the performances of existing works in [2] showed that the sensitivity could score between 82.5% to 100% and specificity between 63.65% to 91.12%, referring mostly to the detection of melanotic lesions against nevus. We also evaluate our experiment using this criterion. Because this criterion is traditionally designed for a two class

		Diagnostic						
		AK	BCC	ML	SCC	SK	Number	Rate
True	AK	8	1	29	7	3	48	16.67%
	BCC	0	229	10	0	27	256	89.45%
	ML	1	6	184	10	8	209	88.04%
	SCC	5	1	38	37	7	88	42.05%
	SK	0	33	11	3	154	201	76.62%
Overall accuracy: OCR(ACR)							812	75.37%(61.89%)

Table 5.4: Confusion matrix for 2D feature set

		Diagnostic						
		AK	BCC	ML	SCC	SK	Number	Rate
True	AK	10	4	28	2	4	48	20.83%
	BCC	1	241	7	0	17	256	94.14%
	ML	0	8	184	14	3	209	88.04%
	SCC	4	3	25	54	2	88	61.36%
	SK	0	21	12	3	165	201	82.09%
Overall accuracy: OCR(ACR)							812	80.54%(68.59%)

Table 5.5: Confusion matrix for 2D+3D feature set

problem, to make it fit our problem, we calculate them for each class individually by treating all other class samples as true negatives. The overall criteria are thereby the mean of the individual results. The results are given in *Table.* 5.6. The average specificities for both s_1 and s_2 are over 90%, which is comparable to most binary classifications reported in the literature (like 74.1% in [39], 92.34% in [10]). On the other hand, a 2% improvement of s_2 over s_1 shows that adding the extra depth information lowers the Type I error (or false positive). Unfortunately, the average specificities are not as good as those reported (like 85.9% in [39], 93.33% in [10]). It is mainly held back by the groups with less samples, such as AK and SCC. This indicates a high false negative rate for these two classes. Though, with the addition of depth features, the average sensitivity is increased from 62% to 69%.

The class specific results are also shown in *Table.* 5.4, *Table.* 5.5 and *Table.* 5.6. For class ML, the two classifiers perform almost the same. Only the specificity is improved very slightly (3%) when taking into account the depth. Typically, ML is a spot with browny colour and regular round or oval shape. These characters make the ML comparatively easy to be distinguished from other lesions by only using colour features. However, from the 3D shape point of view, ML may be flat or it may be raised and its surface can be smooth or rough. Hence, the 3D

	Colour					
Diagnosis	AK	BCC	ML	SCC	SK	Average
Specificity	0.99	0.92	0.85	0.97	0.93	0.93
Sensitivity	0.17	0.86	0.88	0.42	0.77	0.62

	Colour + Depth					
Diagnosis	AK	BCC	ML	SCC	SK	Average
Specificity	0.99	0.93	0.88	0.97	0.96	0.95
Sensitivity	0.21	0.91	0.88	0.61	0.82	0.69

Table 5.6: Classification results evaluated using specificity and sensitivity metrics

features we implement so far make limited contribution to the diagnosis. For all other four classes, whose samples are mostly non-pigmented, the advantage of combining 2D and 3D is obvious. For instance, the improvement to SK between using s_1 and s_2 is 16% (according to OCR). SK often resembles warts and has a rough surface which looks like the scab from a healing wound. This characteristic can be represented by 3D texture features, which help to distinguish them from other lesions. For the dangerous SCC class, including depth-based features outperforms using only colour features by a considerable 20% increase. Commonly, this lesion has a 'crater like' appearance with raised surroundings and a central depression [56]. For them, the eight features utilising depth may be at least partially helpful in representing these characteristics. Even with the large improvement, it still displays a low individual class accuracy and sensitivity on SCC, and similar findings hold true for the AK class. This could be explained by the lack of samples in these two class, without which the classifier cannot be well trained in the learning phase.

2. **Significance Testing**

We test whether the difference in results is statistically significant using McNemar's test on SVM (based on FS II and DB II) results. McNemar's test essentially is based on a χ^2 test and computes a goodness of fit that compares the distribution of counts expected under the null hypothesis (the two systems have the same classification rate) to the observed [56]. The test is based on a 2×2 contingency table, which tabulates the outcomes of two tests. Let n_{10} be the number of examples misclassified by s_2 but not by s_1 and n_{01} be the number of examples misclassified by s_1 but not by s_2. In our case, n_{01} is 82 and n_{10} is 40. The Chi squared statistic equals 13.779 with 1 degree of freedom. The two-tailed P value equals 0.0002. By conventional criteria (0.05 confidence level), this difference is considered to be extremely statistically significant. We also apply the McNemar's test on individual classes. Their respective two-tailed P values are 0.72(AK), 0.05(BCC), 0.81(ML), 0.004(SCC) and 0.07(SK). Under 0.10 confidence level, the improvements in BCC, SCC and SK are statistically significant. Though there is not much difference between the two feature sets for AK and ML.

3. **Comparison between Six Classification Models**

   The performance of six classifiers are compared in *Fig.* 5.5. SVM produces the best result (82.38%) in our application. The confirms the survey result in [2] that the lesion diagnostic systems employing SVM achieved higher performance. The Bayes Classifier using LDA model (81.57%) and QDA model (79.95%) comes the second and third. These three methods clearly outperform the other three, though there is no large difference within them. The reason that LDA is better than QDA might because that the former allows to include more features. KNN comes fourth, with an accuracy of 74.17%. Too much simplification reduces the discriminative power of the Bayes Classifer, the results of DLDA and DQDA are only 69.95% and 69.92%.

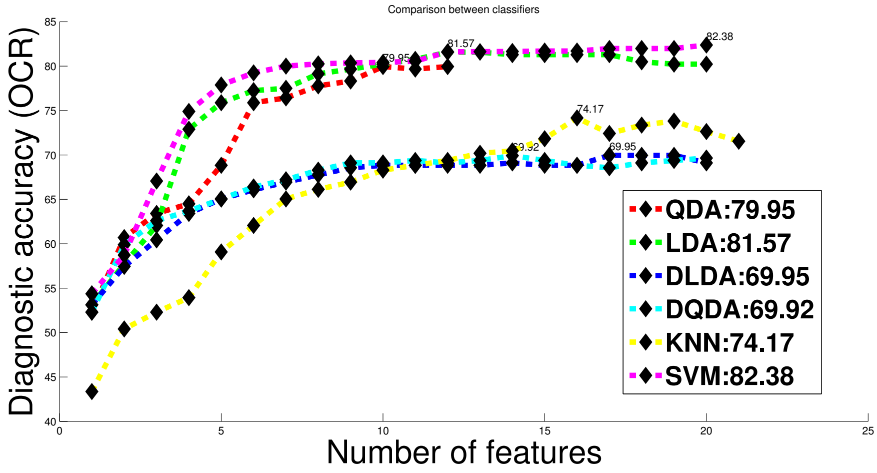


Figure 5.5: Classifier comparison

4. **Comparison between SFFS and SFPS**

   Fig. 5.6 demonstrates the classification results based on feature subsets selected using two strategies: SFFS and SFPS. The Bayes Classifier with the LDA model is used for comparison as it gives a good classification result and is efficient.

The figure shows that SFPS could find a better feature subset for solving the classification problem than SFFS. This improvement might be because 1) SPFS enlarges the search space since it looks for an optimal pair of features rather than a single one in the backward step, 2) similar to the SFFS, SFPS has the ability of re-selecting / removing of the previously discarded / selected features (*e.g.*, in the Feature Set I of DQDA, the first chosen feature 31 is replaced by other features in the latter selection (see *Table.* 5.3)). Although the computational expense increases, it is within an acceptable tolerance. All our experiments are based on the features selected using SFPS. For each classifier model, its feature subset is shown in *Table.* 5.3.

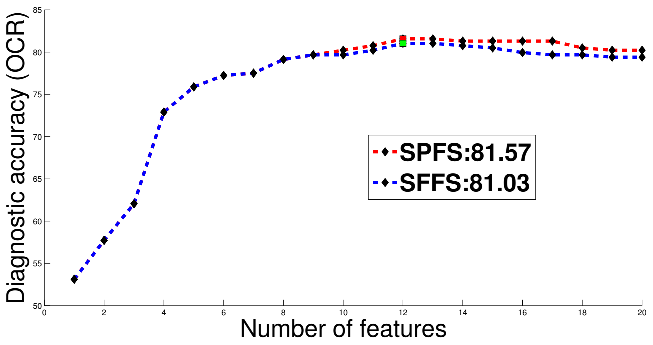


Figure 5.6: Feature selection strategy comparison

5. **Comparison with Human Diagnosis** The computer based classification is surprisingly better than the human (inexperienced medical students) performance. This may be because the computer is able to identify subtle variations in skin which is hard for humans due to the limitation of the naked eye. The SVM classifier produces an accuracy of 82.38% when using 2D and 3D features (FS II) on a database including 812 samples (DB II); while the diagnostic accuracy of the student is only 47% using 3D images and 38.5% using 2D images. Our human raters are medical students who had only undertaken one week of specific dermatological training. Their poor performances reflect the multiple lesion classification task is not trivial for human. Although this comparison is not very fair as they are not based on exactly the same database. The data used for human testing is a sub-set of DATABASE II (including 100 samples) selected by a random strategy. The number of samples in each classifier was identical. But the encouraging performance of the implemented recognition algorithms suggests

such a system can be considered a valuable tool to assist non-experts or even experts when making a judgement.

6. **Important Features for Lesion Diagnosis**
   Table. 5.7 lists the times that each feature is selected by classifiers for the task of classification. The most frequent features (highlighted in red) are 9, 21 and 33. They represent the diagnostic properties of colour variation, local texture and flatness of data distribution, respectively. Their common property is that they are associated with the blue colour channel. This again reflects that skin lesions are more prominent in blue colour [69]. The second important features are highlighted in blue, including two 3D based features, 43 and 45, representing local surface texture and relative flatness of depth data distribution. The skewness related features and the 3D based skin lesion volume are not useful according to the fact that they are rarely selected by classifiers. But in general, various kinds of features (*e.g.,* colour variation based, local texture based, 2D and 3D shape based) have been at least selected once. It is not clear which kind of features are more informative since the feature set selected by different classifiers do not highly agree.

5.6 Conclusion

In clinical settings, in addition to the visual observation of skin lesions, touching is also a part of examination for dermatologists when making a diagnosis. Motivated by this, we hypothesize that both colour and surface shape have discriminative power in lesion diagnosis system. Unfortunately, the state of the art imaging system, dermoscopy, can only provide colour-based information about the skin lesion. Without access to surface shape data, the potential discriminative power of this information has been neglected. New sensors enable simultaneous acquisition of 3D shape and 2D colour information of skin at reasonable resolutions. We first carried out an extensive and rigorous empirical test on medical students to evaluate whether 3D data carried some message resembling the information obtained through touching in the clinic and would help humans in making a correct decision. The diagnostic results support our claim and show a significant increment of 8.5% by using the 3D images. Several auxiliary experiments suggest that this improvement is purely because of the advantage of extra shape information observed from the 3D data rather than any side effects. We further investigate the contribution of 3D data on a computer based five non-melanoma skin lesion classification task by a comparison between using colour features only and using both colour and depth features. We implemented six classifier models as the comparison tool. For each of them, a novel feature selection method (SPFS) that integrated greedy forward and backward strategies is employed to find a locally optimal feature subset that produce the maximal performances. Based on the comparison results, the key point to make is that *the addition of depth does improve the diagnostic accuracy*. This conclusion is found across all six classifiers and under different evaluation metrics (*e.g.,* OCR, ACR, sensitivity and specificity). We also repeat the experiment on a larger database (DATABASE II) using the Support Vector Machine (SVM) which is found to suit our classification task best. For the second time, it confirms that adding the 3D-based features gives an improved classification rate of 80.67% compared to simply colour features (75.25%). The statistical test shows this improvement is significant at the 0.05 confidence level. In addition, we also compare the performances of human and automated computer-based algorithms. The large improvement of the latter suggests that the computer based algorithms might play a valuable role in providing decision-making assistance.

The novelties of our research are 1) we are the first group that has ever extensively studied the importance of 3D data applied to skin lesion classification, 2) we perform the comparison between using 2D data (colour images) and both 2D and 3D data (3D images) on both human and automated computer-based diagnosis, 3) given that most studies reported in the literature focus on the more limited problem of discrimination of melanoma from benign pigmented lesions (mostly moles) [21], we expand the lesion diagnosis study to a broader range of lesions and furthermore, most of these lesions are rarely considered in other works despite their high presence in clinics, 4) we propose a novel feature selection method - SPFS which is shown to outperform SFFS and 5) we compare the performances of six classification models and find the best one for skin lesion diagnosis application.

Given that our ultimate goal is to gain an insight into the benefit of incorporating 3D information rather than pursuing perfect classification performance, we did not put much effort on extracting sophisticated features and enlarging the feature pool. Our feature set only includes 34 colour and 14 depth based features (as listed in *Table.* 5.7). They are mostly basic low-level features. Further improvements would involve developing better features, *e.g.,* ones that model specific characters of different lesions, like the blood vessels of BCC, the cauliflower surface texture of SK, *etc.* These pattern-based features have attracted more attention recently. For example, dermoscopic features - granular, white and hypo-pigmented areas are extracted for the detection of early stage melanoma [28]. In addition, a variety of morphological, clinical and molecular variables could also be taken into account.

Another drawback of the current work lies in the feature selection part. The feature subset of each classifier is chosen by optimizing the Diagnostic Accuracy criterion. This criterion is also used later as a test figure of merit to evaluate the performance of these classifiers. In other words, the comparison between different classifiers is based on the their best performances obtained through adjusting feature subset size and content. Therefore, there is potential risk of data overfitting problem.The features selected in one database may perform badly in another database. As a result, the conclusion obtained in one database may not hold in another database. To assess whether this problem exists, we apply the feature subsets selected under DATABASE I to DATABASE II (see *Table.* 5.2). Fortunately, the same conclusions stay. However, a more proper solution to the problem is to use different criteria for feature selection and performance estimation. Another solution should be dividing the dataset into training and testing sets and selecting the optimal feature subset using the training dataset while evaluating the diagnostic performances using the testing set. The reason that we cannot split the dataset into independent training and testing dataset is the shortage of samples in certain classes (*e.g.,* AK and SCC). This also means that the classifier of certain classes cannot be properly trained. For these classes, their specific classification rates are not ideal. In order to obtain a good variety of these classes, huge numbers of training samples in all classes are required. This may lead to heavy burden to data collection and processing. An alternative way is to only increase the proportion of small classes in the training set and then to compensate the prior probability problem as discussed in 5.2.4. Furthermore, the usefulness of the depth information in the task of classification was investigated only by inexperienced medical students, future work will continue the investigation for experienced dermatologists.

Index	Name	Scheme	Information	# selected	Notes
1	avgColRR	Relative colour brightness features	Colour	4	$\frac{\mu_{a,S}}{\mu_{b,L}}$, S and L denote skin and lesion, respectively
2	avgColGR			4	
3	avgColBR			3	
4	avgColRG			1	
5	avgColGG			5	
6	avgColBG			4	
7	avgColRB			1	
8	avgColGB			3	
9	avgColBB			6	
10	avgRoughR	Relative variability features	Colour	2	$\frac{\sigma_{c,S}}{\sigma_{c,L}}$
11	avgRoughG			3	
12	avgRoughB			3	
36	***avgRoughZ***		**Depth**	3	
13	redtext05	Peak and pit density features	Colour	1	scale = 0.5
14	greentext05			1	
15	bluetext05			2	
41	***ztext05***		**Depth**	3	
16	redtext1		Colour	1	scale = 1
17	greentext1			1	
18	bluetext1			4	
42	***ztext1***		**Depth**	2	
19	redtext2		Colour	1	scale = 2
20	greentext2			2	
21	bluetext2			6	
43	***ztext2***		**Depth**	4	
22	spotDiamDelta20_50	Shape regularity	2D shape	4	
23	KurtosisLesionR	Absolute kurtosis	Colour	5	Flatness of the data distribution
27	KurtosisLesionG			2	
31	KurtosisLesionB			2	
44	***KurtosisLesionZ***		**Depth**	3	
24	SkewnessLesionR	Absolute skewness	Colour	0	Symmetry of the data distribution
28	SkewnessLesionG			1	
32	SkewnessLesionB			0	
46	***SkewnessLesionZ***		**Depth**	1	
25	KurtosisRatioR	Relative kurtosis	Colour	1	Flatness of the data distribution
29	KurtosisRatioG			1	
33	KurtosisRatioB			6	
45	***KurtosisRatioZ***		**Depth**	4	
26	SkewnessRatioR	Relative skewness	Colour	0	Symmetry of the data distribution
30	SkewnessRatioG			0	
34	SkewnessRatioB			1	
47	***SkewnessRatioZ***		**Depth**	2	
35	Area	Area	2D shape	3	
37	***dheight***	Height		2	
38	***i1***	3D moments	**Depth**	2	
39	***i2***			2	
40	***i3***			3	
48	***Volume***	Volume	**Depth**	0	

Table 5.7: List of extracted features. The features derived from 3D data are highlighted in ***bold and italic***. The most frequently selected features are highlighted in red and the second important features are highlighted in green. Feature descriptions can be found in Section 5.4.1

Chapter 6

Conclusion

This thesis aims at investigating the potential benefit of 3D information to multiple class skin lesion diagnosis. A general schematic diagnostic framework comprising data collection and pre-processing, lesion segmentation and classification has been built to explore the contribution of 3D data, particularly on lesion detection and recognition. The preceding chapters present three main contributions:

1. A novel ground truth estimation approach that takes into account the inter-rater variation caused by different diagnostic policies through incorporating a prior pattern analysis of manual segmentation results.

2. A comprehensive evaluation of the contribution of various information on lesion segmentation, particularly focusing on depth information, based on a carefully validated lesion database.

3. A thorough empirical investigation of the contribution of 3D data to both human and computer based non-melanoma skin lesion diagnosis tasks.

These contributions are summarised in the reminder of this chapter, together with a discussion of their limitations and future work.

6.1 Ground Truth (GT) Estimation for Segmentation Evaluation

Main Findings

In order to evaluate our segmentation results derived from Chapter 4, we need a **GT** that is estimated from a collection of manual results. In our work, ground truth estimation is treated as an optimization problem and it is solved under a level-set framework. Three approaches derived from different energy functions are proposed. **LSV** is generated by minimizing the variation between manual segmentations. **LSML** is based on maximizing the *a posteriori* probability (MAP) given a set of manual segmentations and it takes forward the idea in **STAPLE** [1] that takes each rater's performance level into account. A further analysis of manual segmentations reveals that the segmentations of lesions differ mainly because of the rater's segmentation policies. In order to take into account these characteristics of the raters' segmentations, the third approach called **LSMLP** is proposed. **LSMLP** adds an extra energy term related to a Shape Prior Model (SPM) to the energy function of **LSML**. SPM is learned through a prior manual segmentation pattern analysis. Experiments on both synthetic data and real lesion data reveal that **LSMLP** outperforms all the other methods that do not consider the prior information, followed by **LSML** and the state of the art method **STAPLE**. In the field of ground truth estimation, little research has analyzed the patterns of the manual segmentation results and we are the first group that study this subject and integrate it into a ground truth estimation formulation. In addition, we prove theoretically (Section 3.4.1) and experimentally (Section 3.5.1) that **LSV** and *Majority Vote Rule* (**MV**), which produces the smallest average discrepancy between the estimated **GT** and the manual segmentations when a voting threshold is set as $\theta = J/2$ are essentially equivalent.

Limitations and Future Work

1. Our Shape Prior Model (SPM) is generated by combining the **detailed** segmentations through a *Majority Vote Rule* based strategy. Also, in this preliminary study, SPM has a discrete binary formation. It is worth learning the shape prior model in a more comprehensive way, *e.g.,* based on principal components analysis (PCA) or combining the **detailed** segmentations using a better method like **STAPLE** [1] and representing it in a better formulation, *e.g.,* in a continuous format.

2. In order to solve the problem in a level set framework, a simplification is made by assuming that pixels have a spatial independence. To relax this strong assumption, one could introduce a Markov random field model as future work. This might provide better performance on the task of ground truth estimation.

3. Our algorithm can only solve binary segmentation problems. More work needs to be done to extend it to generalized multiple segmentation phase applications. This should

be feasible under a level set framework. In fact, this is the reason that we chose the level set to solve the problem.

6.2 Depth Data Based Lesion Segmentation

Main Findings

For the purpose of comparing the discriminative ability of various lesion information on separating lesion from normal skin, we present a flat segmentation algorithm that is implemented within a level-set framework. The approach is built upon a region-based **deformable model** under a probabilistic formulation. For each lesion region (*i.e.,* lesion and healthy skin), its regional statistical parameters are estimated and used to characterize the respective lesion structures (*i.e.,* lesion and skin) with the density distribution model. We derive and compare several density models. The **GMM** is found to be the best way to model the complex density distribution of properties of skin lesions. The well-known Chan-Vese segmentation [85] is found to be equivalent to the most simplified version in the statistical regional-based segmentation. In order to take into account the prior knowledge of spatial relationships, we introduce the local spatial dependency term, which is modeled by MRF in the form of a Gibbs distribution.

The segmentation property comparison results reveal that integrating depth and texture properties results in a significant overall improvement of segmentation, in terms of both accuracy and consistency. By adding depth and texture properties, the error rate is reduced from $7.80\% \pm 5.35\%$ to $6.62\% \pm 2.60\%$. This convincing gain suggests that the depth and texture descriptors are complementary to colour. From this respect, the depth information does improve lesion segmentation. On the other hand, we find that colour properties enable a close segmentation to the dermatologists on pigmented lesions and the depth and texture information is particularly meaningful for non-pigmented lesions which have less colour variation over different regions. This suggests that different information has specific importance for different types (*e.g.,* pigmented and non-pigmented) of lesions and only discriminative and useful properties should be chosen for each lesion according to their appearances, instead of using all the properties derived from all kinds of information for every lesion. Hence, we further propose a hierarchical segmentation structure that incorporates pattern recognition techniques to the flat segmentation. The novel segmentation strategy performs segmentation by using different feature subsets within a hierarchical structure which is in turn determined by the discrimination between different lesion appearances. The experiments show that both the one layer hierarchical approach based on colour and the two layer hierarchical approach based on colour and depth further improve the segmentation results compared to the flat segmentation approach (from 6.62% to 6.15%).

In our study, we also included several non-pigmented skin lesion types. These lesion types have been rarely discussed before in the medical imaging community. Adding them increases the difficulty of segmentation tasks because they are comparatively lacking the most important discriminative cue (*i.e.,* colour).

Limitations and Future Work

Our hierarchical structure is built up in an *ad-hoc* way according to some kinds of prior information. In the future, a more sophisticated hierarchical structure could be built based on 1) a better definition of the hierarchical layer splitting parameters and 2) additional lesion data with ground truth segmentations that allow more extensive development and testing.

6.3 Depth Data Based Lesion Diagnosis

Main Findings

The empirical tests using medical students reveals that viewing 3D lesion data carries some message resembling the information obtained through touching the lesion in the clinic and would help humans in making a correct decision. Compared to the 2D images, the human-based diagnostic results show a significant increase of 8.5% by using the 3D images. Several auxiliary experiments suggest that this improvement is purely because of the advantage of extra shape information observed from the 3D data rather than any side effects.

A further comparison between using colour features only and using both colour and depth features shows that the 3D data also improves the computer based classification performance over the five non-melanoma skin lesion classes. This conclusion holds over the six classifier models that are implemented and the two databases that are researched on. Through comparison, we find the Support Vector Machine (SVM) outperforms the other classifiers in the lesion classification task. An experiment using SVM on DATABASE II shows that adding 3D-based features gives an improved classification rate of 80.67% compared to simply colour features (75.25%). This result is extremely significant under the 0.05 confidence level.

For the purpose of finding better feature subsets, we propose a new feature selection method, SPFS, which incorporates forward and backward greedy selection in a novel strategy. The comparison shows that SPFS could select a better sub-optimal feature subset that SFFS.

Limitations and Future Work

In our work, we did not concentrate our effort on extracting sophisticated features and enlarging the feature pool, because improving the diagnosis using better features is not the goal of this research. The features used in the thesis are mostly from the preliminary study in our group carried out by McDonagh *et al.* [56]. Though, in the future, improvement on the diagnostic system would involve developing better features, *e.g.,* ones that model specific characters of different lesions, like the blood vessel of BCC, the cauliflower surface texture for SK, *etc.* These pattern-based features have attracted more attention recently. For example, the dermoscopic features - granular, white and hypo-pigmented areas are extracted for the detection of early stage melanoma [28]. In addition, a variety of morphological, clinical and molecular variables, such as the age of the patient, lesion location, *etc.*, should also be taken into account.

Because of the shortage of samples in certain classes (*e.g.,* AK and SCC), our databases are not balanced. The diagnosis result may have been biased to the large lesion groups. For building a classifier, collecting enough images is an important issue to ensure system accuracy and generality [20]. Otherwise, for the classes with insufficient samples, as the classifier model cannot be properly trained, their specific classification rates are not ideal. Adding more samples to these classes seems to be very necessary. Besides, to solve the unbalanced database problem, another solution is to further divide the lesion class with large samples into sub-groups, given the finding that there exist sub-classes for certain lesions (*e.g.,* BCC and melanoma) which are 'almost' discriminant based on the image features or histological features. This is a brand new topic and has been rarely investigated, except for a recent work conducted by Armengol [119] who introduced a method called *LazyCL* for generating a domain theory to classify melanomas. We believe incorporating this sub-classification scheme would benefit the diagnostic system with better classifier models that represent classes with significantly different appearances.

In summary, this thesis raises a scientific question of whether there is potential benefit of the 3D information to the computer aided skin lesion diagnosis given the fact that viewing the lesion superficial shape is also a part of lesion diagnosis in clinic. The extensive experiments based on a complete computer-based diagnostic system have given a positive answer to the question. We have proven that the 3D depth data of lesions embeds useful messages about the lesion location (Section 4) and identity (Section 5). We suggest that by adding 3D to the colour information which is commonly used in the skin lesion diagnostic community, there is a chance of improving the diagnostic rate of computer-based diagnostic systems. This work provides evidence for the future lesion diagnostic systems to take advantage of lesion surface shape information. But it must be noted that, at the current stage, the benefit from the 3D data is small (*e.g.,* the improvement is 1.2% for segmentation and 5% for classification) and the use of depth data is not always possible due to stereo failures. Furthermore, the 3D data capturing system is expensive. However, with the fast development of 3D data capturing system and the improvement of 3D data analysis algorithms, there is hope to fulfill the ultimate goal in the skin lesion analysis field of providing objective, consistent, quantitative, cheap and accurate automatic diagnostic results.

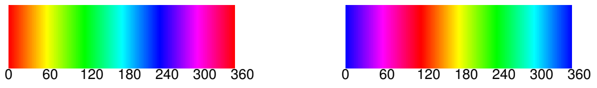
Appendix A

Hue Modification Using a Shifted Scale

As it can be seen in *Fig.* A.1(a), the hue colour representation is arranged around the colour wheel, which starts from red at 0° and wraps back to red at 360°. The histogram analysis shows that the hue of lesion images is narrowly distributed at the red colour region, as shown in *Fig.* A.2(d) and (e). The majority of the hue values centralize around the narrow interval between 0° and 60°. For some particular lesions, there are also some values located near 360°, which also belong to the red colour category. In such cases, the hue feature cannot be directly used for representing the pure colour property of lesions because of the break of the red colour region (see *Fig.* A.2(b)). To solve the problem, we modify the hue value by a simple shifting operation. The goal is to shift the colour circle break point to another colour. We choose to use blue, which is placed at $H = 240°$, as it is not a valid colour for lesions. Therefore, the colour transformation has the form as:

$$\widehat{H} \equiv \begin{cases} H - 240°, H \geq 240° \\ H + 121°, H < 240° \end{cases} . \tag{A.1}$$

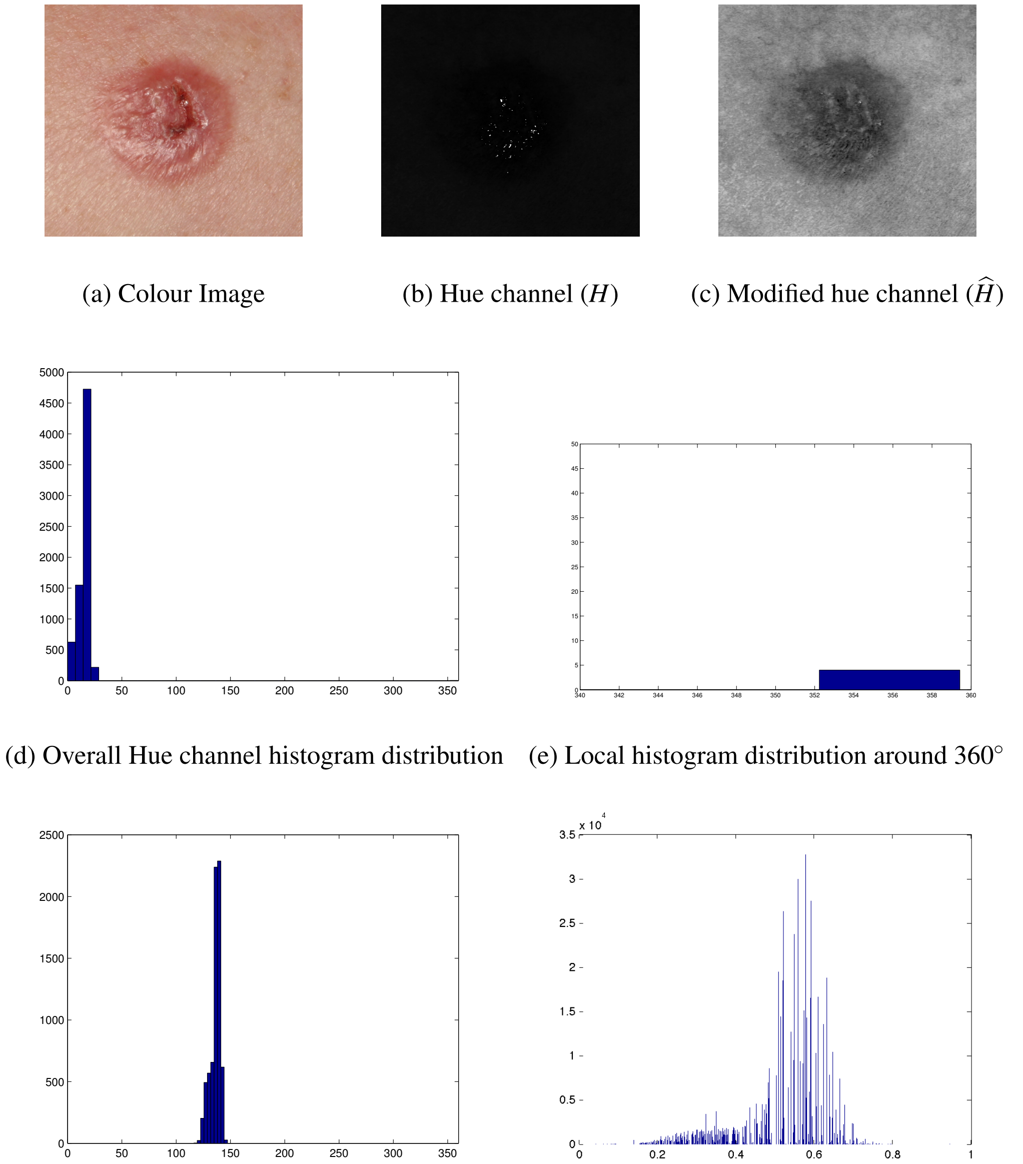
As shown in *Fig.* A.1(b), after the modification, the hue still has the range from 0° to 360°, but the break of the colour wheel takes place at blue colour. Red is placed at



(a) Hue

(b) Modified Hue

Figure A.1: Hue colour wheel.



(a) Colour Image

(b) Hue channel (H)

(c) Modified hue channel ($\widehat{H}$)

(d) Overall Hue channel histogram distribution

(e) Local histogram distribution around 360°

(f) Hue channel histogram distribution after the modification

(g) Hue channel histogram distribution after the projection

Figure A.2: Histogram distribution of the hue channel of lesion case D489.

$H = 120°$. Now, pixels with red colour are guaranteed to be assigned to continuous values around 120°, as shown in *Fig.* A.2(f). We further project the hue value to the interval $[0,1]$ using the following transformation:

$$\widehat{H} = \frac{\widehat{H} - \min \widehat{H}}{\max \widehat{H} - \min \widehat{H}}. \tag{A.2}$$

The histogram distribution of the projected $\widehat{H}$ is given in *Fig.* A.2(g). From *Fig.* A.2(c), one can see that the modified hue channel can properly present the pure colour property of lesion case D489.

This hue channel transformation has been performed on all the lesion data. A visual inspection of the transformed hue images shows that this method solves the colour wheel breaking problem.

Appendix B

Stereo System Accuracy Test

Evaluating stereo system accuracy is difficult, because of the lack of a proper testing target which should have a known micro-wise scale and rich texture. To get a general idea of the capturing ability of our system, we conduct a preliminary experiment. Our testing targets are two crossing surgical sutures that are tightly attached onto a flat surface (as shown in the left column of *Fig.* B.1). The goal is therefore to find the distance between two sutures when they can no longer be separated on the depth image. The experimental steps are given as following:

First, draw a line across the two sutures, which is shown as the blue vertical bar in *Fig.* B.1. Its 3D profile is shown on the right column in *Fig.* B.1. The line then moves in the horizontal direction towards the right gradually and automatically.

Second, find the intersecting points between the sutures and the blue vertical bar. The point detection is based on the colour image using a thresholding technique. The middle column in *Fig.* B.1 shows this result. The background points are assigned to 0 and the intersecting points to 1.

Third, project those detected 2D points onto the 3D space. The associated 3D points are denoted as pt_{i1}, pt_{j2}, where $i \in 1, \ldots, I$ and $j \in 1, \ldots, J$ indicate the index of points on suture 1 and 2, respectively. They are highlighted in red on the depth profiles, where two peaks could be spotted. They correspond to the sutures and the detected points located around them. When two peaks exist and are distinguishable, it means that the stereo system is able to detect and separate them. The distance between two points is calculated as $\frac{\sum_{i,j} ||pt_{i1} - pt_{j2}||}{I \times J}$.

As the blue line approaches the intersection points of the two sutures gradually, the peaks tend to merge, as shown in the bottom row of *Fig.* B.1. It indicates that the two sutures could no longer be separated when the distance between textures is less than

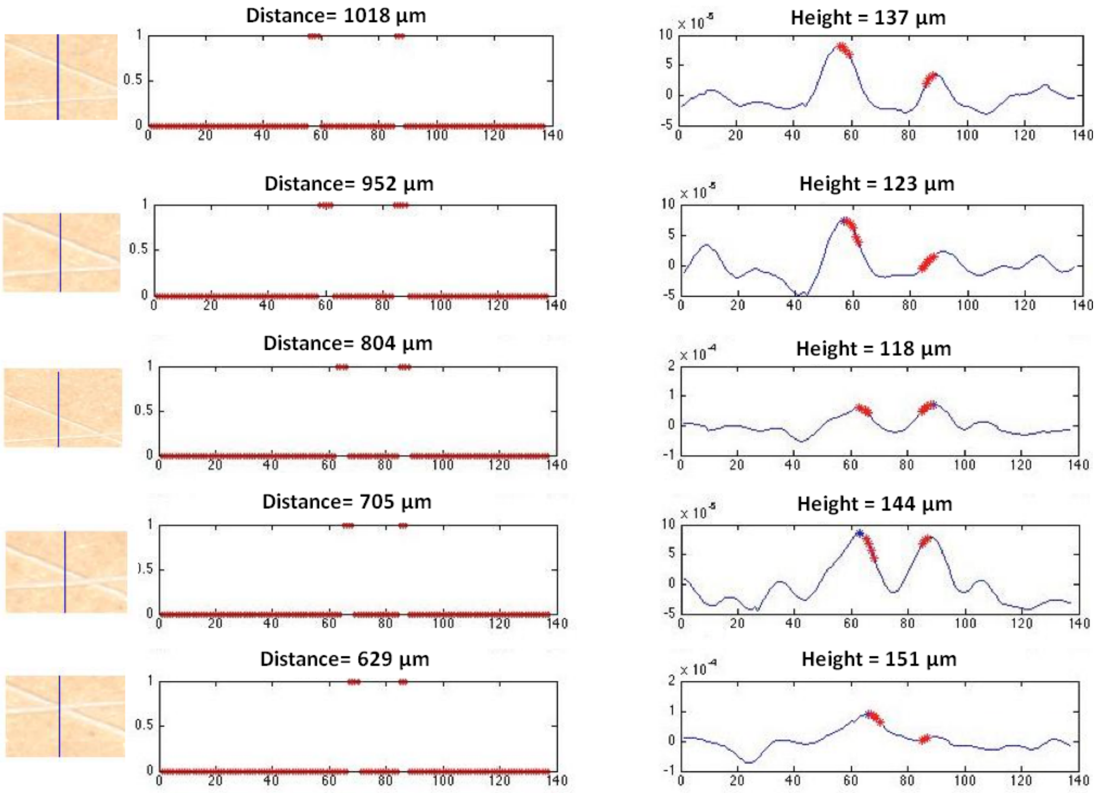


Figure B.1: Stereo system test diagram: the blue vertical bar approaches the intersecting point of two sutures in the horizontal direction by a small step setup by the algorithm.

$629\mu m$. Hence, we estimate the texture separating capability of the system should be about $705\mu m$, under which the sutures could still be separated. On the other hand, in order to evaluate the system accuracy in capturing fine textures, we also measure the diameter of the suture. This parameter is estimated as the average height of the top points of sutures based on the 3D profile. The height of the peaks is calculated as the difference between the peak point and a line representing the flat background surface, which is fitted using the points along the 3D profile excluding the points in the local neighborhood of the peaks. Based on five measurements, the suture diameter is estimated as $134.6\mu m(\pm 13.9)$. For the purpose of system evaluation, the physical measurement of the suture is taken as the ground truth. The measurement was conducted by Craig Walker in Dermatology Department, Edinburgh University and is based on an average of ten measurements. The result is $187.9\mu m$. The difference between the 3D and physical measurements is $53.3\mu m$. This large discrepancy might be partly caused by the inaccurate estimation of the suture height in 3D data, because it is very difficult to find a flat surface which is also capturable using the 3D system. In our experiment, the flat surface is obtained by binding a paper with rich texture onto a flat glass surface. Previously we measured the depth RMS for our system as $25\mu m$ and the inter pixel separation as $30\mu m$. This means that the texture scale must be on the order of $629/30\mu m \simeq 30$ pixels. The result shows that the 3D system has not matured to a point to accurately capture very fine details of objects. However, one must see that the 3D system is able to **detect** and **separate** fine textures with $0.7mm$ scale. Hence, there is no doubt that they could be used to capture lesion surface variation.

Appendix C

Pseudocode of Segmentation Methods

Flat segmentation in pseudocode

```
Initialize lesion boundary
WHILE iteration number is less than N (or algorithm does not converge)
    Estimate the pdfs inside and outside lesion boundary ($p_1/p_2$)
    FOR each pixel along the boundary ($x$)
        Calculate Eq. 4.33
        Adjust the level set function ($\phi(x)$)
    ENDFOR
    Test for convergence
ENDWHILE
```

Segmentation in hierarchical structure pseudocode

```
Initialize lesion boundary
   Calculate DV using C
   Compare DV with a pre-set threshold
   IF the lesion belongs to pigmented lesion category
      Perform Flat segmentation using C
   ELSE
      Calculate DV using D
      Compare DV with a pre-set threshold
      IF the lesion belongs to non-flat lesion category
         Perform Flat segmentation using CD
      ELSE
         Perform Flat segmentation using CDT
```

C denotes colour-based features

CD denotes colour and depth based features

CDT denotes colour, depth and texture based features

DV denotes the regional dissimilarity value between lesion and skin.

Appendix D

Image Database

Collecting sufficient lesion images is an important issue for building a classifier and to ensure system accuracy and generality [20]. This is not a trivial task, because a good database needs large amount of high quality samples with diagnosis information and have been pre-processed into a usable format. All these lead to heavy workload.

Our skin lesion data pool has been continuously expanded over the past four years because of the ongoing collection in the Dermatology Department of Edinburgh University. The number of samples reached 2001 in April, 2010. The imaging equipment is the Dimensional Imaging [19] dense stereo image capture system which allows simultaneous acquisition of 3D shape and colour data of skin. The system is built around a pair of Canon EOS 350D SLR cameras. Each camera acquires a 3456x2304 image. Given camera placement, lenses and patient placement, each pixel corresponds to about 0.03 mm skin sample separation. Measurements have determined an RMS depth error also of about 0.03 mm. More information on the system capabilities can be found in Appendix B.

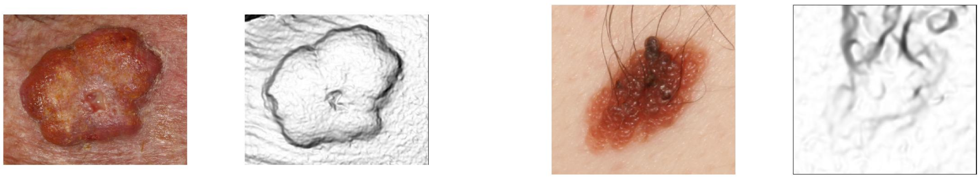
Regarding factors like the sample amount and 3D reconstruction quality, only limited number of samples are kept. The selection is based on a visual inspection of the cosine-projection (cosine shading) of depth image and colour image (see *Fig.* D.1). Samples are excluded if: diagnosis is ambiguous, depth recovery failed or colour image is unsatisfactory. As a result, two lesion databases (Database I and Database II) have been installed and applied in succession in our work. Database I and Database II have 369 and 812 samples, respectively. All the samples come from five lesion classes, which are AK (Actinic keratosis), BCC (Basal cell carcinoma), SCC (Squamous cell carcinoma), ML (Melanocytic nevus) and SK (Seborrheic keratosis). Their distributions are shown in *Table.* D.1, respectively. The deadly form of skin cancer - melanoma is not included because of the shortage of samples. AK and SCC classes are comparatively underrepresented in both databases because of the limited patient availability. Another reason for the shortage of AK samples is the poor 3D reconstruction. AK often occurs on the top of the head of aged people, where lots of the hair noises result in catastrophic problems for the 3D building.

		Lesion					
		AK	BCC	ML	SCC	SK	Total Number
DATABASE I	Number	14	140	83	17	115	369
	Percentage	3.79%	37.94%	22.49%	4.61%	31.17%	
DATABASE II	Number	48	256	209	88	201	812
	Percentage	5.99%	31.92%	26.06%	10.97%	25.06%	

Table D.1: Lesion distributions in DATABASE I and DATABASE II

In addition, for the convenience of further analysis, some pre-processing steps are necessary: 1) cropping which removes the background region (*e.g.,* the capturing frame) and enables the following operations to focus on the lesion area (see *Fig.* D.2). 2) rotation of the 3D data that makes sure the lesion's surface is fronto-parallel (see *Fig.* 4.3) and 3) isolating the lesion from the normal skin region prior to the classification step using the segmentation algorithm developed in Chapter 4.

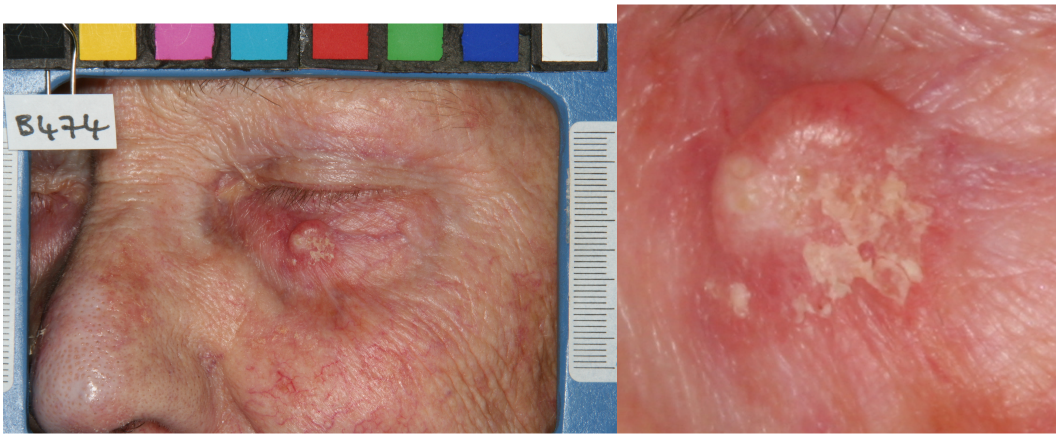
To obtain the diagnosis information, dermatologists usually need histopathological tests or long term clinical follow-up [20]. In our project, the diagnosis of the lesions is established based on opinion from multiple dermatologists' clinical observation and histopathology and it is considered as the Ground Truth in our supervised classification.



(a) Good 3D reconstruction

(b) Bad 3D reconstruction

Figure D.1: Good and bad lesion data examples observed on the cosine-projection of depth image(right) and colour image(left)



(a) Full data

(b) Cropped data

Figure D.2: An example of the cropped lesion data

Appendix E

Publications

The research presented in this thesis has also formed the basis for several peer-reviewed papers, listed as follows:

1. Li X and Aldridge B and Ballerin L and Fisher RB and Rees J. Estimating the ground truth from multiple individual segmentations incorporating prior pattern analysis with application to skin lesion segmentation. *International Symposium on Biomedical Imaging (ISBI)*, pages 1438-1441, 2011.

2. Li X and Aldridge B and Ballerin L and Fisher RB and Rees J. Estimating the ground truth from multiple individual segmentations with application to skin lesion segmentation. *Medical Image Understanding and Analysis (MIUA)*, 1(1):101-106, 2010.

3. Li X and Aldridge B and Ballerin L and Fisher RB and Rees J. Depth data improves skin lesion segmentation. *In Medical Image Computing and Computer-Assisted Intervention MICCAI 2009 12th International Conference*, volume 12, pages 1100-1107, 2009

4. Aldridge B, Li X, Ballerin L, R. Fisher RB, Jonathan L. Rees, Teaching Dermatology Using 3-Dimensional Virtual Reality, Correspondence, *Archives of Dermatology*, 146(10), Oct 2010.

5. Ballerini L, Li X, Fisher RB, Aldridge B, Rees J, Content-Based Image Retrieval of Skin Lesions by Evolutionary Feature Synthesis, *Proceeding of the 12th European Workshop on Evolutionary Computation in Image Analysis and Signal Processing*, Istanbul, pages 312-319, April 2010.

6. Ballerini L, Li X, Fisher RB, Aldridge B, Rees J, A Query-by-Example Content-Based Image Retrieval System of Non-Melanoma Skin Lesions, *Proceeding of MICCAI-09 Workshop MCBR-CDS 2009: Medical Content-based Retrieval for Clinical Decision Support*, London, Caputo B *et al.*. (Eds.): MCBR_CBS 2009, LNCS 5853, pages 31-38. Springer-Verlag, Heidelberg, 2010.

Bibliography

[1] Warfield SK, Zou KH, and Wells WM. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. *IEEE Transactions on Medical Imaging*, 23(7):903–921, 2004.

[2] Maglogiannis I and Doukas CN. Overview of advanced computer vision systems for skin lines characterization. *IEEE Transaction on Information Technology Biomedicine*, 13(5):721–33, 2009.

[3] Skin cancer (non-melanoma), December 12, 2010.

[4] Cascinelli N, Ferrario M, Tonelli T, and Leo E. A possible new tool for clinical diagnosis of melanoma: The computer. *Journal of the American Academy of Dermatology*, 16(2):361–367, 1987.

[5] Taouil K and Romdhane NB. Automatic segmentation and classification of skin lesion images. *The 2nd International Conference on Distributed Frameworks for Multimedia Applications*, 1(1):1–12, 2006.

[6] Day GR and Barbour RH. Automated melanoma diagnosis: where are we at? *Skin Research and Technology*, 6(1):1–5, 2000.

[7] Piccolo D, Ferrari A, Peris K, Diadone R, Ruggeri B, and Chimenti S. Dermoscopic diagnosis by a trained clinician *vs.* a clinician with minimal dermoscopy training *vs.* computer-aided diagnosis of 341 pigmented skin lesions: a comparative study. *British Journal of Dermatology*, 147(3):481–486, 2002.

[8] Claridge E, Cotton S, Moncrieff M, and Hall P. *Chapter 37, Spectrophotometric intracutaneous imaging (SIAscopy): method and clinical applications. Handbook of non-Invasive methods and the skin, Second Edition.* CRC Press, 2006.

[9] Cotton S. A noninvasive skin imaging system. *Technical Report, CSR-97-3. University of Birmingham School of Computer Science*, 1997.

[10] Celebi ME, Kingravi HA, Uddin B, Iyatomi H, Asl YA, Ogan A, Stoecker WV, and Moss RH. A methodological approach to the classification of dermoscopy images. *Computerized Medical Imaging and Graphics*, 31(6):362–373, 2007.

[11] Ding Y, Smith L, Smith, Sun JA, and Warr R. Obtaining malignant melanoma indicators through statistical analysis of 3D skin surface disruptions. *Skin Research and Technology*, 15(3):262–270, 2009.

[12] Cyganek B and Siebert JP. *An introduction to 3D computer vision techniques and algorithms*. Wiley, 2007.

[13] Igarashi T, Nishino K, and Nayar SK. The appearance of human skin. *Technical Report, Department of Computer Science, Columbia University CUCS-024-05*, 2005.

[14] Leveque JL. EEMCO guidance for the assessment of skin topography. *Journal of the European Academy of Dermatology and Venereology*, 12(2):103–114, 1999.

[15] Manousaki AG, Manios AG, Tsompanaki EI, Panayiotides JG, Tsiftsis DD, Kostaki AK, and Tosca AD. A simple digital image processing system to aid in melanoma diagnosis in an everyday melanocytic skin lesion unit: a preliminary report. *International Journal of Dermatology*, 45(4):402–410, 2006.

[16] Jacobi U, Chen M, Frankowski G, Sinkgraven R, Hund M, Rzany B, Sterry W, and Lademann J. *In vivo* determination of skin surface topography using an optical 3D device. *Skin Research and Technology*, 10(4):207–214, 2004.

[17] Sun J, Smith M, Smith L, Midha S, and Bamber J. Object surface recovery using a multi-light photometric stereo technique for non-lambertian surfaces subject to shadows and specularities. *Image and Vision Computing*, 25(7):1050–1057, 2007.

[18] Bennamoun M and Mamic GJ. *Object recognition: fundamentals and case studies*. Springer, 2002.

[19] Dimensional Imaging - world leading 3D and 4D surface imaging, March 2011.

[20] Iyatomi H. *Computer-based diagnosis of pigmented skin lesions, New Developments in Biomedical Engineering*. InTech, 2010.

[21] Stoecker WV, Wronkiewiecz M, Chowdhury R, Stanley RJ, Xu J, Bangert A, Shrestha B, Calcara DA, Rabinovitz HS, Oliviero M, Ahmed F, Perry LA, and Drugge R. Detection of granularity in dermoscopy images of malignant melanoma using colour and texture features. *Computerized Medical Imaging Graphics*, 35(2):144–147, 2011.

[22] Li X, Aldridge B, Ballerin L, Fisher RB, and Rees J. Depth data improves skin lesion segmentation. In *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2009*, volume 12, pages 1100–1107, 2009.

[23] Li X, Aldridge B, Rees J, and Fisher RB. Estimating the ground truth from multiple individual segmentations incorporating prior pattern analysis with application to skin lesion segmentation. *IEEE International Symposium on Biomedical Imaging (ISBI)*, 1(1):1438–1441, 2011.

[24] Li X, Aldridge B, Rees J, and Fisher RB. Estimating the ground truth from multiple individual segmentations with application to skin lesion segmentation. *Medical Image Understanding and Analysis (MIUA)*, 1(1):101–106, 2010.

[25] Green A, Martin N, McKenzie G, Pfitzner J, Quintarelli F, Thomas BW, O'Rourke M, and Knight N. Computer image analysis of pigmented skin lesions. *Melanoma Research*, 1(4):231–236, 1991.

[26] She Z and Excell PS. Skin pattern analysis for lesion classification using local isotropy. *Skin Research and Technology*, 17(2):206–212, 2011.

[27] Celebi ME, Stoecker WV, and Moss RH. Advances in skin cancer image analysis. *Computerized Medical Imaging Graphics*, 35(2):83–84, 2011.

[28] Dalal A, Moss RH, Stanley RJ, Stoecker WV, Gupta K, Calcara DA, Xu J, Shrestha B, Drugge R, Malters JM, and Perry LA. Concentric decile segmentation of white and hypopigmented areas in dermoscopy images of skin lesions allows discrimination of malignant melanoma. *Computerized Medical Imaging and Graphics*, 35(2):148–154, 2010.

[29] Baldi A, Quartulli M, Murace R, Dragonetti E, Manganaro M, Guerra O, and Bizzi S. Automated dermoscopy image analysis of pigmented skin lesions. *Cancers*, 2(2):262–273, 2010.

[30] Iyatomi H, Oka H, Celebi ME, Hashimoto M, Hagiwara M, Tanaka M, and Ogawa K. An improved internet-based melanoma screening system with dermatologist-like tumor area extraction algorithm. *Computerized Medical Imaging and Graphics*, 32(7):566–579, 2008.

[31] Garnavi R, Aldeen M, Celebi ME, Varigos G, and Finch S. Border detection in dermoscopy images using hybrid thresholding on optimized colour channels. *Computerized Medical Imaging and Graphics*, 35(2):105–150, 2011.

[32] Capdehourat G, Corez A, Bazzano A, and Muse P. Pigmented skin lesions classification using dermatoscopic images. *Proceedings of the 14th Iberoamerican Conference on Pattern Recognition: Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications (CIARP)*, 1(1):537–544, 2009.

[33] Stolz W, Holzel D, and Riemann A. Multivariate analysis of criteria given by dermoscopy for the recognition of melanocytic lesions. *Book of Abstracts, Fiftieth Meeting of the American Academy of Dermatology*, 1991.

[34] Derma medical.

[35] Schaefer G, Rajab MI, Celebi ME, and Iyatomi H. Colour and contrast enhancement for improved skin lesion segmentation. *Computerized Medical Imaging and Graphics*, 35(2):99–104, 2011.

[36] Iyatomi H, Celebi ME, Schaefer G, and Tanaka M. Automated colour calibration method for dermoscopy images. *Computerized Medical Imaging Graphics*, 35(2):89–98, 2010.

[37] Application of photometric stereo in dermatology, October 04, 2010.

[38] Ng V and Cheung D. Measuring asymmetries of skin lesions. *IEEE International Conference on Systems, Man, and Cybernetics Conference Proceedings*, 5:4211–4216, 1997.

[39] Grana C, Pellacani G, Cucchiara R, and Seidenari S. A new algorithm for border description of polarized light surface microscopic images of pigmented skin lesions. *IEEE Transactions on Medical Imaging*, 22(8):959 – 964, 2003.

[40] Lee TK and Claridge E. Predictive power of irregular border shapes for malignant melanomas. *Skin Research and Technology*, 11(1):1–8, 2005.

[41] Ganster H, Pinz A, Rohrer R, Wildling E, Binder M, and Kittler H. Automated melanoma recognition. *IEEE Transactions on Medical Imaging*, 20(3), 2001.

[42] Stanley RJ, Stoecker WV, and Moss RH. A relative colour approach to colour discrimination for malignant melanoma detection in dermoscopy. *Skin Research and Technology*, 13(1):62–72, 2007.

[43] Casari A, Pellacani G, Seidenari S, Cesinaro AM, Beretti F, Pepe P, and Longo C. Pigmented nodular basal cell carcinomas in differential diagnosis with nodular melanomas: confocal microscopy as a reliable tool for *in vivo* histologic diagnosis. *Journal of Skin Cancer*, 1(1):1–7, 2011.

[44] Fabbrocini G, Betta G, Leo GD, Liguori C, Paolillo A, Pietrosanto A, Sommella P, Rescigno O, Cacciapuoti S, Pastore F, Vita VD, Mordente I, and Ayala F. Epiluminescence image processing for melanocytic skin lesion diagnosis based on 7-point check-list: A preliminary discussion on three parameters. *The Open Dermatology Journal*, 4:110–115, 2010.

[45] Betta G, Leo GD, Fabbrocini G, Paolillo A, and Sommella P. Dermoscopic image-analysis system: estimation of atypical pigment network and atypical vascular pattern. *IEEE International Workshop on Medical Measurement and Applications*, 1(1):63–37, 2006.

[46] Celebi ME, Iyatomi H, Stoecker WV, Moss RH, Rabinovitz HS, Argenziano G, and Soyer HP. Automatic detection of blue-white veil and related structures in dermoscopy images. *Computerized Medical Imaging and Graphics*, 32(8):670–677, 2008.

[47] Zhou Y, Smith M, Smith L, Farooq A, and Warr R. Enhanced 3D curvature pattern and melanoma diagnosis. *Computerized Medical Imaging and Graphics*, 35(2):155–165, 2010.

[48] Menzies SW, Bischof L, Talbot H, Gutenev A, Avramidis M, Wong L, Lo SK, Mackellar G, Skladnev V, McCarthy W, Kelly J, Cranney B, Lye P, Rabinovitz H, Oliviero M, Blum A, Varol A, De'Ambrosis B, McCleod R, Koga H, Grin C, Braun R, and Johr R. The performance of SolarScan: an automated dermoscopy image analysis instrument for the diagnosis of primary melanoma. *Archives of Dermatology*, 141:1388–1396, 2005.

[49] Iyatomi H, Oka H, Celebi ME, Ogawa K, Argenziano G, Soyer HP, Koga H, Saida T, Ohara K, and Tanaka M. Computer-based classification of dermoscopy images of melanocytic lesions on acral volar skin. *Journal of Investigative Dermatology*, 128(8):2049–2054, 2008.

[50] Ballerini L, Li X, Fisher RB, and Rees J. A query-by-example content-based image retrieval system of non-melanoma skin lesions. *MICCAI09 Workshop: Medical Content-based Retrieval for Clinical Decision Support, London, B. Caputo et al. (Eds.): MCBR_CBS 2009, LNCS 5853*, 1(1):31–38, 2009.

[51] Rubegni P, Cevenini G, Burroni M, Perotti R, Dell'eva G, Sbano P, Miracco C, Luzi P, Tosi P, Barbini P, and Andreassi L. Automated diagnosis of pigmented skin lesions. *International Journal of Cancer*, 101(1):576–580, 2002.

[52] Iyatomi H, Norton KA, Celebi M, Schaefer G, Tanaka M, and Ogawa K. Classification of melanocytic skin lesions from non-melanocytic lesions. *IEEE Proceedings of Engineering Medicine and Biology Society*, 1(1):5407–5410, 2010.

[53] Caslellini P, Scalise A, and Scalise L. A 3D measurement system for the extraction of diagnostic parameters in suspected skin nevoid lesions. *IEEE Transactions on Instrumentation and Measurement*, 49(5):924–928, 2000.

[54] Callieri M, Cignoni P, Pingi P, Scopigno R, and Coluccia M. Derma: monitoring the evolution of skin lesions with a 3D system. *Vision, Modeling and Visualization Workshop (VMV)*, pages 19–21, 2003.

[55] Leveque JL and Querleux B. SkinChip, a new tool for investigating the skin surface *in vivo*. *Skin Research and Technology*, 9(1):343–347, 2003.

[56] McDonagh S, Fisher RB, and Rees J. Using 3D information for classification of non-melanoma skin lesions. *Medical Image Understanding and Analysis (MIUA)*, 1:164–168, 2008.

[57] Prastawa M, Bullitt E, and Gerig G. Synthetic ground truth for validation of brain tumor MRI segmentation. *Medical Image Computing and Computer-assisted Intervention (MICCAI)*, 3749:26–33, 2005.

[58] Warfield SK, Zou KH, and Wells WM. Validation of image segmentation by estimating rater bias and variance. *Phil. Trans. R. Soc. A*, 366(1874):2361–2375, July 2008.

[59] Joel G, Schmid-Saugeon P, Guggisberg D, Cerottini JP, Braun R, Krischer J, Saurat JH, and Murat K. Validation of segmentation techniques for digital dermoscopy. *Skin Research and Technology*, 8(4):240–249, 2002.

[60] Kittler J, Hatef M, Duin RPW, and Matas J. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239, 1998.

[61] Commowick O and Warfield SK. Estimation of inferential uncertainty in assessing expert segmentation performance from staple. *IEEE Transactions on Medical Imaging*, 29(3):771–780, 2010.

[62] Langerak TR, van der Heide UA, Lips IM, Kotte ANTJ, Vulpen MV, and Pluim JPW. Label fusion using performance estimation with iterative label selection. *IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1186–1189, 2009.

[63] Klein S, van der Heide UA, Lips IM, van Vulpen M, Staring M, and Pluim JP. Automatic segmentation of the prostate in 3D MR images by atlas matching using localized mutual information. *Medical Physics*, 35(4):1407–1417, 2008.

[64] Celebi ME, Schaefer G, Iyatomi H, Stoecker WV, Malters JM, and Grichnik JM. An improved objective evaluation measure for border detection in dermoscopy images. *Skin Research and Technology*, 15(4):444–450, 2009.

[65] Heckemann RA, Hajnal JV, Aljabar P, Rueckert D, and Hammers A. Automatic anatomical brain MRI segmentation combining label propagation and decision fusion. *NeuroImage*, 33(1):115–126, 2006.

[66] Chabrier S, Laurent H, Emile B, Rosenberger C, and Marche P. A comparative study of supervised evaluation criteria for image segmentations. *The 14th European Signal Processing Conference (EUSIPCO 2006)*, 1(1):1143–1146, 2006.

[67] Osher S and Sethian JA. Fronts propagating with curvature-dependent speed: Algorithms based on hamilton-jacobi formulations. *Journal of Computational Physics*, 79(1):12–49, 1988.

[68] Puranik P, Bajaj P, Abraham A, Palsodkar P, and Deshmukh A. Human perception-based colour image segmentation using comprehensive learning particle swarm optimization. *The 2nd International Conference on Emerging Trends in Engineering and Technology (ICETET), 2009*, 1(1):630–635, 2009.

[69] Celebi ME, Iyatomi H, Schaefer G, and Stoecker WV. Lesion border detection in dermoscopy images. *Computerized Medical Imaging Graphics*, 33(2):148–153, 2009.

[70] Pantofaru C and Hebert M. A comparison of image segmentation algorithms. *Technical Report, CMU-RI-TR-05-40, The Robotics Institute, Carnegie Mellon University*, 2005.

[71] Ma Z, Tavares JM, Jorge RN, and Mascarenhas T. A review of algorithms for medical image segmentation and their applications to the female pelvic cavity. *Computer Methods in Biomechanics and Biomedical Engineering*, 13(2):235–246, 2010.

[72] Darrell S. Rigel, Julie Russak, and Robert Friedman. The evolution of melanoma diagnosis: 25 years beyond the ABCDs. *CA-Cancer Journal for Clinicians*, 60:301–316, 2010.

[73] Xu L, Jackowski M, Goshtasby A, Roseman D, Bines S, Yu C, Dhawan A, and Huntley A. Segmentation of skin cancer images. *Image and Vision Computing*, 17(1):65–74, 1999.

[74] Glasbey CA and Horgan GW. *Image analysis*. John Wiley and Sons, Ltd, 2006.

[75] Kass M, Witkin A, and Terzopoulos D. Snakes: Active contour models. *International Journal of Computer Vision*, 1(4):321–331, 1988.

[76] Tang J. A multi-direction GVF snake for the segmentation of skin cancer images. *Pattern Recognition*, 42(6):1172–1179, 2009.

[77] Zhou H, Schaefer G, Celebi ME, Lin F, and Liu T. Gradient vector flow with mean shift for skin lesion segmentation. *Computerized Medical Imaging Graphics*, 35(2):121–127, 2010.

[78] Lankton S and Tannenbaum A. Localizing region-based active contours. *IEEE Transactions on Image Processing*, 17(11):2029–2039, 2008.

[79] Deng HW and Clausi DA. Unsupervised image segmentation using a simple MRF model with a new implementation scheme. *International Conference on Pattern Recognition (ICPR)*, 2:691–694, 2004.

[80] Wang H, Moss RH, Chen X, Stanley RJ, Stoecker WV, Celebi ME, Malters JM, Grichnik JM, Marghoob AA, Rabinovitz HS, Menzies SW, and Szalapski TM. Modified watershed technique and post-processing for segmentation of skin lesions in dermoscopy images. *Computerized Medical Imaging Graphics*, 35(2):116–120, 2010.

[81] Iyatomi H, Oka H, Saito M, Miyaka A, Kimoto M, Yamagami J, Kobayashi S, Tanikawa A, Hagiwara M, Ogawa K, Argenziano G, Soyer HP, and Tanaka M. Quantitative assessment of tumour area extraction from dermoscopy images and evaluation of the computer-based methods for automatic melanoma diagnostic system. *Melanoma Research*, 16(2):183–190, 2006.

[82] Yuan X, Situ N, and Zouridakis G. A narrow band graph partitioning method for skin lesion segmentation. *Pattern Recognition*, 42(6):1017–1028, 2009.

[83] Chung DH and Sapiro G. Segmenting skin lesions with partial-differential-equations-based image processing algorithms. *IEEE Transactions on Medical Imaging*, 19(7):763–767, 2000.

[84] Erkol B, Moss R, Stanley R, Stoecker W, and Hvatum E. Automatic lesion boundary detection in dermoscopy images using gradient vector flow snakes. *Skin Research and Technology*, 11(1):17–26, 2005.

[85] Chan TF and Vese LA. Active contours without edges. *IEEE Transactions in Image Processing*, 10(2):266–277, 2001.

[86] Cremers D, Rousson M, and Deriche R. A review of statistical approaches to level set segmentation: integrating colour, texture, motion and shape. *International Journal of Computer Vision*, 72(2):195–215, 2007.

[87] Brox T and Weickert J. Level set based image segmentation with multiple regions. *In Pattern Recognition, Springer LNCS 3175*, 1(1):415–423, 2004.

[88] Mete M, Kockara S, and Aydin K. Fast density-based lesion detection in dermoscopy images. *Computerized Medical Imaging Graphics*, 35(2):128–136, 2011.

[89] Sande KEA, Gevers T, and Snoek CGM. Evaluating colour descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1582–1596, 2010.

[90] Woods K. Genetic algorithms: colour image segmentation literature review. 2007.

[91] Fischer S, Schmid P, and Guillod J. Analysis of skin lesions with pigmented networks. *International Conference on Image Processing*, 1:323–326, 1996.

[92] Brox T and Weickert J. A TV flow based local scale measure for texture discrimination. *Proceedings of the 8th European Conference on Computer Vision (ECCV)*, 2:578–590, 2004.

[93] Xie XH. Level set based segmentation using local feature distribution. *Internal Conference on Pattern Recognition (ICPR)*, pages 2780–2783, 2010.

[94] Dhawan AP and Sim A. Segmentation of images of skin lesions using colour and texture information of surface pigmentation. *Computerized Medical Imaging and Graphics*, 16(3):163–177, 1992.

[95] Freeman WT and Adelson EH. The design and use of steerable filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(9):891–906, 1991.

[96] Perona P and Malik J. Scale-space and edge detection using anisotropic diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(7):629–639, 1990.

[97] Weeratunga SK and Kamath C. PDE-based non-linear diffusion techniques for denoising scientific and industrial images: an empirical study. *Image Processing: Algorithms and Systems Conference, SPIE Electronic Imaging Symposium*, 4667:279–290, 2002.

[98] Dalal N and Triggs B. Histograms of oriented gradients for human detection. *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2(1):886–893, 2005.

[99] Ott P and Everingham M. Implicit colour segmentation features for pedestrian and object detection. *IEEE 12th International Conference on Computer Vision (ICCV)*, pages 723–730, 2009.

[100] Ma J and He Q. A dynamic merge-or-split learning algorithm on gaussian mixture for automated model selection. *Intelligent Data Engineering and Automated Learning - IDEAL 2005*, 3578:203–210, 2005.

[101] Tiplica T, Verron S, Kobi A, and Nastac I. FDI in multivariate process with naive bayesian network in the space of discriminant factors. *International Conference on Computational Inteligence for Modelling Control and Automation and International Conference on Intelligent Agents Web Technologies and International Commerce (CIMCA'06)*, 1(1):216–221, 2006.

[102] Peserico E and Silletti A. Is (N)PRI suitable for evaluating automated segmentation of cutaneous lesions? *Pattern Recognition Letters*, 31(16):2464–2467, 2010.

[103] Nachbar F, Stolz W, Merkle T, and Cognetta AB. The ABCD rule of dermatoscopy: high prospective value in the diagnosis of doubtful melanocytic skin lesions. *Journal of the American Academy of Dermatology*, 30:551–559, 1994.

[104] Henry GI, Grevious MA, Morton TA, and Stadelmann WK. Skin, benign skin lesions, March 2011.

[105] Sigurdsson S, Hansen LK, and Drzewiecki K. Colour segmentation of skin lesions with the generalizable gaussian mixture model. *IMM Technical report 2003-22*, 2003.

[106] Azofra AA, Aznarte JL, and Benitez JM. Empirical study of feature selection methods based on individual feature evaluation for classification problems. *Expert Systems with Applications*, 38(7):8170–8177, 2011.

[107] She Z, Duller AWG, and Fish PJ. Enhancement of lesion classification using divergence, curl and curvature of skin pattern. *Skin Research and Technology*, 10(4):222–230, 2004.

[108] Iyatomi H, Oka H, Tanaka M, and Ogawa K. Parametric analysis of acral lesions on dermoscopy. *International Conference on Complex Medical Engineering*, 1(1):336–339, 2007.

[109] Talavera L. An evaluation of filter and wrapper methods for feature selection in categorical clustering. *Advances in Intelligent Data Analysis VI, Lecture Notes in Computer Science*, 3646:440–451, 2005.

[110] Burrell LS, Smart OL, Georgoulas G, Marsh E, and Vachtsevanos GJ. Evaluation of feature selection techniques for analysis of functional MRI and EEG. *International Conference on Data Mining*, 1(1):256–262, 2007.

[111] Aly M. Survey on multiclass classification methods. Technical report, California Institute of Technology, 2005.

[112] Bishop CM. *Neural networks for pattern recognition*, volume Section 6.5. Oxford University Press, USA; 1 edition, 1995.

[113] Hsu CW and Lin CJ. A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks*, 13(2):415–425, 2002.

[114] Dreiseitl S, Ohno-Machado L, Kittler H, Vinterbo S, Billhardt H, and Binder M. A comparison of machine learning methods for the diagnosis of pigmented skin lesions. *Journal of Biomedical Informatics*, 34(1):28–36, 2001.

[115] McDonagh S. Skin cancer surface shape based classification. *Thesis, School of Informatics, University of Edinburgh*, 2008.

[116] Huang DS, Quan Y, He M, and Zhou BS. Comparison of linear discriminant analysis methods for the classification of cancer based on gene expression data. *Journal of Experimental and Clinical Cancer Research*, 28(1):149, 2009.

[117] Sidney S. Nonparametric statistics for the behavioral sciences. *New York: McGraw-Hill*, pages 75–83, 1956.

[118] Chang CC and Lin CJ. *LIBSVM: a library for support vector machines*. National Taiwan University, http://www.csie.ntu.edu.tw/ cjlin/libsvm, 2001.

[119] Armengol E. Classification of melanomas *in situ* using knowledge discovery with explained case-based reasoning. *Artificial Intelligence in Medicine*, 51(2):93–105, 2011.