University of Edinburgh
School of Informatics

Skin Cancer Surface Shape Based Classification

4th Year Project Report
Artificial Intelligence and Computer Science

Steven McDonagh (0458953)

March 4, 2008

Abstract: The classification of skin cancer typically relies on colour and intensity images as input to a variety of rule and checklist based diagnosis. Computer vision techniques can be applied to enhance such image data into representative features, in order to then automate the detection process by machine learning methods. In general, these approaches do not take into account any information that could additionally be provided by depth and the resulting three dimensional surface structure description of the skin area in question.

A dense stereo photogrammetry technique is used to gain both colour and depth based information from skin lesions. A system is developed to investigate the usefulness of depth information as an additional modality for enhanced dermatological classification. The available channels of information are converted and processed to enhance features, distributions within the data are investigated and classification experiments are performed with a Bayesian classifier delivering classification accuracies of 77.3% 83.7%.

Acknowledgements

I would like to express my gratitude to my supervisor, Professor Bob Fisher, whose expertise, support, attention to detail and hard work helped me throughout the course of the project.

In addition, I would like to thank Professor Jonathan Rees for providing the project data samples and offering his valuable expert opinion.

Contents

1. Introduction

The classification of skin cancer has conventionally relied on colour and intensity images as input to a variety of rule and checklist based diagnosis. Various computer vision techniques can be applied to enhance such image data into more representative features, in order to then automate the detection process by pattern recognition and machine learning methods.

In general, these approaches do not take into account any information that could additionally be provided by *depth* and the resulting three dimensional *surface structure* description of the skin area in question.

In this project we investigate whether the inclusion of depth based information can infact help to distinguish between various types of skin cancer and other dermatological lesion classes. We aim to show that the actual skin surface structure may provide a valuable source of features on which to base classification. Our hypotheses relating to this goal are detailed and stated explicitly in Section 1.2.

The project involves the conversion and processing of the available channels of information to enhance features, the analysis of any variations and distributions within the data to help differentiate between classes and evaluating the classification performance that depth based features allow. We use a dense stereo photogrammetry technique to gain both colour images and depth based information in what can be considered a new application of a well known method.

To summarise, the benefits of using surface shape information for automated skin cancer classification are investigated in this project.

1.1 Project overview

The core components of the project involve making use of preprocessing techniques to enhance image data, defining and constructing numerical features that are descriptive of skin images and building a skin image classification system for evaluating the constructed features.

Primarily, we looked at developing and selecting representative features of the collected image data that attempt to take advantage of the available depth based information. This depth based information is provided by our novel data capture technique. Secondly, a classification system was built to assess the usefulness of the selected features.

The intended outcome of the work is to provide evidence that image features which make use of depth based information are indeed useful for the purpose of skin cancer classification.

1.2 Hypotheses of the project

We now explicitly state the claims made in our work and describe the evidence that must be provided to support them. The primary goal of the project is to investigate any additional benefits that depth information might provide over standard image data for skin cancer classification. This aim can be encapsulated by the following testable statement which includes a prediction.

1.2.1 Main hypothesis

A classification system using a combination of standard and depth based image features is more successful at the task of classifying skin lesion images than a system which uses standard image features alone.

The testing of this hypothesis will involve comparing our baseline classification system (using only standard image features as defined in Section 2.2) with a system that makes use of additional depth and surface structure based information when classifying skin images. The principal dimensions on which these systems are to be compared are that of classification accuracy and *cost* (these metrics are defined formally in Sections 7.1.2.1 and 7.1.2.2). The hypothesis can therefore be considered a claim about the proposed systems along the scientific behavioural dimension of correctness.

The required evidence for this claim is experimental in nature. We aim to produce image classification results that exhibit a significantly better classification accuracy or minimum cost for the system that makes use of both depth and standard features compared to the baseline system which makes use of standard features alone.

1.3 Structure of the document

To achieve the proposed goals of the project we first carry out a review of related research in the field of automated skin cancer classification which can be found in Chapter 2. Following this, the data made use of in this project and the equipment needed to provide it are described in Chapter 3.

To present evidence in support of our hypothesis we build a skin lesion classification system capable of preprocessing and extracting features from the provided data. We then define and apply classification rules to derive a classifier from the sample data. The methodology employed by these components of the system and details of their implementation can be found in Chapters 4, 5 and 6.

The results obtained by testing the implemented system are quoted and analysed in Chapter 7 and the concluding Chapter contains discussion and recommendations relating to the work carried out.

2. Related research

There exists a fairly well established body of work related to the problem of automated skin cancer classification. It has been nearly twenty years since the first mention was made of utilising the computer as a tool for assisting clinicians in diagnosing skin lesions [7]. Automated *Melanoma* diagnosis in particular, is a relevant problem which has been extensively addressed by a large body of work and is a well studied problem in the literature [11, 19].

Typically, images of skin lesions are analysed by computer to capture certain features (expressed as numeric values) thought to be characteristic of malignancy. These features can then be used to classify the image and report a diagnosis. An overview of the main techniques and approaches associated with this area of research follows.

2.1 Data capture

2.1.1 Standard image capture techniques

Pre 1995, the most popular method of clinical diagnosis was examination with the naked eye by an experienced clinician. Most computer based methods therefore, also used "clinical view" (naked-eye images) as input for their systems in this early period [20, 36]. Post 1995, what can be considered one of the major developments in the automated diagnosis of skin cancer was the advent of the Epiluminescent Microscopy technique.

Epiluminescent Microscopy (also known as ELM, Dermatoscopy and Skin Surface Microscopy) began to gain popularity as a technique for diagnosing melanoma. This is a noninvasive technique that involves applying a low power television microscope and light to a glass slide covering mineral oil on the surface of a skin lesion. The goal is to make the epidermis translucent, enabling more visual information to be gathered than from naked-eye images alone. From this period onwards, reports about ELM research in this area became more common, and research into ELM based automated classification systems became increasingly popular [8, 34].

The techniques outlined above can be considered the most widely used image capture approaches for the purpose of skin cancer classification and are the standard alternatives to the method used in this project. The main difference between the discussed techniques and those employed in this project is that the standard techniques do not directly capture any *depth* information from the image scene.

2.1.2 Alternative data capture approaches

In the work surveyed, only one other independent system has been identified that attempts to utilise depth data as a modality for image acquisition. This is the DERMA system of Callieri et al [10].

The main difference between the DERMA system and the work proposed here is the technique by which the depth data is acquired. The DERMA system is based on the laser scanning of a subject to obtain 3D and aligned colour information while our work makes use of dense stereo photogrammetry (see Section 3.1 for an outline of our technique).

A further noted difference is that the primary intended use of the DERMA system is to measure and assess the *time evolution* of chronic skin wounds. In contrast, the main purpose of our system is as a tool with which to assess the suitability of depth based image features for lesion classification purposes. For this reason, a direct comparison is deemed an unsuitable measure of the success of this project.

An alternative approach with respect to extracting novel types of information (other than depth based) is explored in [35] where Raman spectra[1] are obtained by pointing a laser beam at the skin sample to be classified. The laser beam excites molecules in the sample and the frequency shifts of the reflected spectra are interpreted as functions of the type of molecules in the sample, thus the Raman spectra can be used to obtain useful information on the different chemical compounds in the examined skin. This feature information is then used with machine learning techniques for classification purposes.

2.2 Feature derivation and image analysis

In the majority of classification systems found in the literature, images are analysed with algorithms derived from criteria provided by experts. The most popular example is the well-known ABCD criteria proposed by Friedman et al [30]. The ABCD rule represents the *asymmetry*, *border* structure, variegated *colour*, and *diameter* of the skin lesion and define the basis for a diagnosis by a dermatologist.

[1] A spectroscopic technique used in condensed matter physics and chemistry to study vibrational modes in a system. Vibrational information is specific to chemical bonds in molecules and provides a fingerprint by which a particular molecule can be identified.

The main reason for the popularity of this criteria appears to be the ease with which it is reproduced algorithmically, rather than its clinical usefulness. There are a number of examples of systems which make use of features derived in this way which report very good results [18, 19].

Part of our work involves reproducing some of these "standard" features derived from regular colour images such that they may then serve as a baseline for comparison with the depth based features that we define (cf. Chapter 5).

Some preliminary work [26, 31] has been carried out with the stereo photogrammetry camera rig made use of in this project. In his Masters Thesis Ravindranath defines a number of pixel colour and depth features (means and variations) which were extracted from the image data and were used with a Bayes classifier to distinguish between lesion classes. Ravindranath's Thesis is summarised in Section 2.4. This previous work can be considered directly related to our project and a basis which can be extended, primarily by exploring the benefit of extracting many more and richer properties from the data. It is in this way that our work can be regarded as building upon existing work in the field.

2.3 Classification technique

During the survey of the related literature, it was noted that classification techniques tend to vary from paper to paper. Example techniques include expert systems, discriminant analysis, decision trees, artificial neural networks, logistic regression and Bayesian learning based methods. A summary of classification techniques used and the results achieved was compiled by Day et al in [17]. This information is reproduced here in Table 2.1.

This lack of coordinated effort and standard classification technique is what Hall et al [20] term "directing computer technology at this problem in an *ad hoc* fashion". The variety of applied classification techniques in the field is a factor that would make direct and scientific comparison to other systems on the strength of our depth based feature set difficult.

2.4 Previous work with 3D depth data

Ravindranath [31] designed and evaluated a skin spot classification system that made use of both intensity and 3D range data captured from the same equipment that is used in this project (see Section 3.1 for hardware specification). Within the surveyed literature, this can be considered the most closely related work to that which was carried out in this project. For this reason, we provide a brief

Author	Year	Segmentation technique	Classifier	Number of algorithms	Sens.	Spec.	Number of images	Image Type
Dhawan	1988	Automatic: partially described	Expert system	18: partially described 20: patient data	Work in progress	Work in progress	Not available	'Nevoscope' images
Green et al.	1991	Automatic: partial description	Discriminant analysis	11: described	80%	72%	70/5	Digitised video images
Cascinelli et al.	1992	Automatic: not described	Thresholding and feature summation	8: partially described	83%	60%	169/45 train, 44/10 test	TV camera mag: ×3
Schindewolf et al.	1992	Automatic: described	Decision trees (CART)	Unknown. described in general	94%	88%	353/215	Digitised slides mag: ×0.5
Bostock et al.	1993	See (26)	Artificial neural network (ANN)	Shape only. described in (26)	92%	68%	124/68 62/37 train 62/31 test	See (26)
Ercal et al.	1994A	Hand drawn	ANN	Exp 1: 14* Exp 2: 8 described	Exp 1a: 80% Exp 1b: 86% Exp 2a: 74% Exp 2b: 82%	Exp 1: 86% Exp 1b: 86% Exp 2a: 84% Exp 2b: 83%	Exp A: 240/120 Exp B: 216/108	Digitised slides
Green et al.	1994	Automatic: partially described	Discriminant analysis	17: described	83%	82%	164/18	CCD video camera
Ercal et al.	1994B	Hand drawn	ANNs and Fuzzy inference	17: described	96%	62%	399/135	Digitised slides
Andreassi et al.	1995	Automatic: not described	Not applicable	8: described	Unknown	Unknown	430/50	Digitised video images
Hintz-Madsen et al.	1996	Unknown	ANN	21: colour and texture not described	59% of the test set was classified correctly	180/60 train, 60 test (unknown distribution)	180/60 train 60 test (unknown distribution)	Clinical view digitised photographs
Menzies et al.	1997	Automatic: not described	Statistical: logistic regression	Unknown: not described, commercial	93%	67%	170/75	Digitised dermaphot
Gutkowicz-Krusin et al.	1997	Automatic: described	Statistical: linear classifier	13: partially described	All 100% *	Exp 1: 92% Exp 2: 92% Exp 3: 61%	140/30 Exp1&2: 76/24 Exp 3: train 76/25 test 28/5 **	Digitised ELM slides - Only MM and AMN
Seidenari et al.	1998	Automatic: not described	Statistical: discriminant analysis	21: described	93%	95%	917/65 used in study 90/31 compared to human diagnosis	Video microscope. Mag: ×10-50
Bischof et al.	1998	Automatic: partially described	Decision tree (RPART)	Unknown: not described, commercial	89-100% †	80-84%	221/45	ELM video camera
Ravindranath ‡	2005	Hand drawn	Naive Bayes	7: described	Exp 1: 50% of the test set was classified correctly	Exp 2: 81.81% - 87.5% of the test set was classified correctly	Exp 2: train: x/66 (unknown distribution) test: 11/18	Digitsed images & 3D range data

* The linear classifier was constrained to detect 100% of melanoma.

** Of the three experiments, only Experiment 3 used a separate training set.

† Re-substitution gave 100%, whilst cross-validation gave 89%. They state the likely value is somewhere between the two.

‡ This work was not present in the summary by Day et al. Included here by the author.

Table 2.1: Summary of related research. Source: [17]

review and discussion of the system designed and evaluated by Ravindranath and the classification results obtained.

The system described was used to extract a total of 17 features from a dataset of 84 images of skin lesions originating from seven distinct classes. The classes of skin lesion studied were Basel Cell Carcinoma, Melanoma, Squamous Cell Carcinoma, Melanocytic Nevus, Dermatofibroma, Seborrhoeic Keratosis and Cherry Nevus. The data set was provided by the Department of Dermatology at the University of Edinburgh. The features that were evaluated included colour metrics, texture metrics and shape metrics which were derived from the intensity (colour) data and a single depth based feature, maximum spot height with respect to the surrounding skin surface, derived from the 3D range data. A second range data feature was proposed but not implemented, the processing time required for 3D feature extraction from the high resolution (3456×2304) image data and time constraints were cited as the reasons for not implementing this second range data based feature. The features that were implemented included means and variation of the RGB and HSV channels of each image, area, eccentricity, perimeter and maximum height values for each data sample.

The variation of the extracted features were analysed across each lesion class and a subset of six features thought to encapsulate the highest discriminatory power were retained and then used with a Naive Bayes Classifier for the purpose of classifying the skin lesions. The full set of 17 features were not used for classification purposes primarily due to the lack of available samples for some of the represented skin lesion classes. For example only six samples were available for the classes 'Melanoma', 'Cherry Nevus' and 'Dermatofibroma'.

To explore the usefulness of the depth based feature for classification, two systems were trained. The first used six intensity features while the second made use of five of these intensity features and substituted the sixth feature that exhibited the least variation within the training samples for the depth based feature. The classifiers were both trained on 66 lesion samples and tested on the remaining 18 samples. Both systems were then evaluated based on classification accuracy. The classification results of these seven class experiments did not differ between the two systems and both classifiers were shown to classify nine of the 18 test images correctly.

On inspection of the classification results provided, it is noted that both systems were able to classify all test samples belonging to Melanocytic Nevus and Basel Cell Carcinoma correctly, and all other test samples were misclassified as one of these two classes. It was proposed by the author that this trend of correctly and incorrectly classified samples may have been due to the fact that the two classes mentioned had the largest amount of training data available, with at least 14 samples from each class in the training set and therefore also the largest *apriori* probabilities.

With both of the trained systems making identical misclassifications and therefore returning the same accuracy results, it was noted that no proof to support the hypothesis that *"the use of depth information may be a significant discriminatory feature for skin spot classification"* could be provided. However, Ravindranath suggests that if there existed sufficient training data for each of the seven classes, the classifier would be able to exhibit an accurate classification performance across all classes. This claim can be partially substantiated by a second experiment that divided the training and testing data samples into just two classes "Benign" and "Malignant", thus providing more training data for each class. In this second experiment, the system was able to achieve an 87.5% accuracy for the class of benign samples and 81.81% for malignant samples.

This project attempts to build on the previous work carried out in [31] primarily by extracting more and richer intensity and depth based properties from the available data set and by evaluating classifiers on accuracy and further metrics. By reducing the resolution of image data (by a cropping technique), we have aimed to partially circumvent the problem of high feature extraction times. Due to the fact that several hundred more data samples have been collected since this work was carried out, we also attempt to explore the achievable classification performance on a realistic data set.

2.5 Summary

The proposed technique of using depth based information for skin image classification can be regarded as a fairly novel approach. Some related previous work with the image capture device has been carried out by Ravindranath. With respect to the surveyed material, only one other example of independent work making use of depth information was found.

As previously noted, the ad hoc direction and reporting of research in this field makes direct system comparison difficult and *"reinvention of the wheel"* within the field likely [17]. A variety of different techniques are used for image capture, image analysis and classification.

In general, image preprocessing techniques tend to be developed specifically for the data set being worked with. There appears to be no evidence of a standard approach and techniques tend not to be fully detailed, making reproduction of research results difficult. The lack of a standard set of test images is a noted difficulty with research in this area.

In contrast to this, during their survey of the field Day et al [17] conclude that feature derivation and image analysis algorithms tend to be well described in general. Unfortunately, in some cases the basic bench mark of the research being replicable is not met. This failure may be due to commercial interests (e.g. Bischof et al. [8]).

3. Data and equipment

3.1 Stereo photogrammetry equipment

The intensity images and depth data is captured via a dense stereo photogrammetry technique. The dense stereo (depth) data is recovered from two simultaneously captured images via stereo photogrammetry matching software which is an extended version of PSP^{TM}technology, an off-the-shelf innovative product from Dimensional Imaging (http://www.di3d.com/). The core technology of the product is summarised in [1]. The basis of this reconstruction is a multi-resolution algorithm, an earlier version of which is described in [2].

This technique has the benefits over previous medical skin research that uses laser scanning (namely the DERMA system discussed in Section 2.1.2) of instantaneous capture and perfect 1:1 alignment of colour information to z-depth. Instantaneous capture does not produce artifacts due to subject movement and the method can be considered non-invasive and completely safe since it uses standard photographic equipment.

The method allows us to capture accurate 3D data at a very fine scale. The data was captured by a stereo capture rig constructed of two 8.0 MPixel Canon EOS 350D cameras, calibrated and using the maximum level of magnification supported by the standard EF-S lens (0.28m closest focus distance). A ring flash is employed for consistent lighting of the subject and a Macbeth colour chart is attached to the perimeter of the capture area. A picture of the stereo camera rig can be found in Figure 3.1.

The capture quality and scale of the depth images[1] have been measured at a z-noise of 0.028 microns deviation from the plane and x,y pixel separation values of 0.033mm and 0.031mm respectively. This enables extremely close-up stereo photogrammetry to capture the micro scale 3D structure present in the skin samples. The depth information represents the z-axis distance from the sensor. These depth values are then used to provide an extra modality for classification and enhanced visualisation.

[1] As measured by X. Li, a PhD student within the IPAB institute in the School of Informatics, The University of Edinburgh.

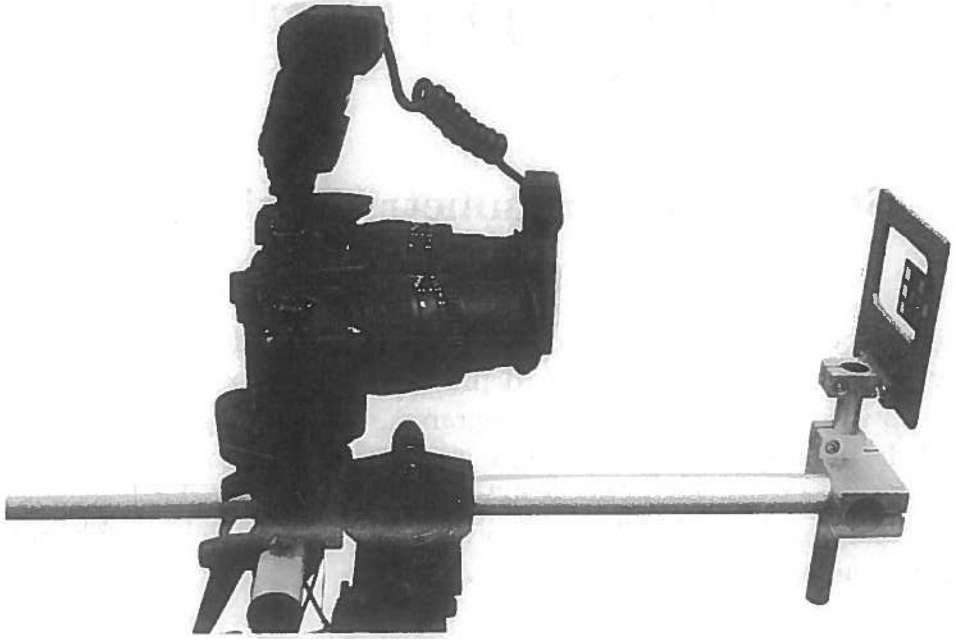


Figure 3.1: Stereo camera rig. Original image due to [37]

3.2 Data description

The resultant data produced by the equipment described in Section 3.1 is in the form of a standard intensity true colour RGB image of each skin lesion and the accompanying range data in the form of matrix data structures containing the z-axis distances of the skin surfaces from the sensor. The original intensity images have a spatial resolution of 3456×2304 pixels with the inter-pixel spacing $\approx 25 pixel/mm$. A Macbeth chart is part of the capture equipment but no normalisation of colour is performed during the imaging process, partly due to the reportedly ineffective results of colour normalisation cited by [31]. Example intensity and depth data captured by the stereo equipment can be seen in Figure 3.3.

The skin lesion data used in this project has been provided by the *Department of Dermatology* at the *University of Edinburgh*. Overall 378 skin lesion images categorised into twelve original classes were available to us. These were categorised as *Actinic keratosis, Basal cell carcinoma, Comedone, Dermatofibroma, Haemagioma, Melanoma, Pyogenic granuloma, Seborrheic keratosis, Squamous cell carcinoma, Benign melanocytic naevus (normal moles), Viral wart and Blue naevus.* All images were captured during clinical examinations by the capture system in a working dermatology clinic. The ground truth for the lesion classes was generated by clinical diagnosis from an expert from the *Department of Der-*

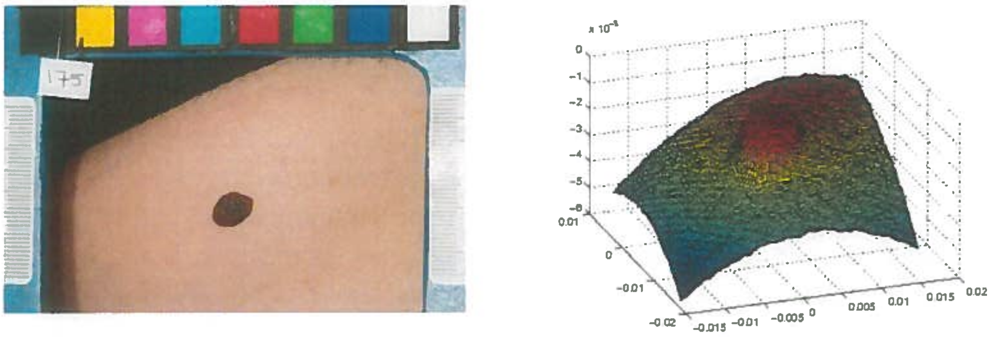


Figure 3.2: Intensity and depth data recovered from a skin lesion sample

matology.

The number of samples per class was not evenly distributed, leaving some of the classes in the original data set with extremely few samples for the purposes of model building and testing (for example there are only two specimens of *Comedone* in the data set and only six *Viral warts*). Five classes of skin lesion from this original set of twelve were deemed to have enough samples available with which to perform meaningful experimentation. These were *Actinic keratosis (AK), Basal cell carcinoma (BCC), Seborrheic keratosis (SK), Squamous cell carcinoma (SCC)* and *Benign melanocytic naevus (ML)* giving a total of 234 usable intensity images and accompanying range data for the purposes of training and testing the system. These five classes were used throughout the remainder of this project and the number of available samples for each class are found in Table 3.1

Lesion class	Number of available samples
AK	11
BCC	65
ML	61
SCC	25
SK	72

Table 3.1: Skin lesion classes and occurrence frequencies

3.3 Visual skin lesion characteristics

Here we give a brief overview of the visual characteristics for each of the five classes of skin lesion used for experimentation in this project. A survey of any notable or distinguishing visual or structural features of each class may help to provide insight into the feature selection (Chapter five) and Classification (Chapter six) stages of the project.

Skin lesions are typically subdivided into Malignant and Benign categories. Malignant is the medical term used to describe a severe and progressively worsening disease and as such, Malignant lesions can be thought of as highly dangerous. Conversely, Benign lesions are non-cancerous and therefore provide significantly lower risk to patients. Of the five classes examined, *Basel cell carcinoma* and *Squamous cell carcinoma* are Malignant growths, *Seborrheic keratosis* and *Normal moles* are regarded as Benign and *Actinic keratosis* is typically classed as a precancerous growth that may turn Malignant in the future [3]. Typical visual characteristics for each of the five classes according to the Skin Cancer Foundation [3] are listed below. Some further visual symptoms listed are due to [31] and originate from the expert opinion of a dermatologist who has collaborated on work related to this project

3.3.1 Basel cell carcinoma

The most common form of skin cancer, Basal Cell Carcinoma is a malignant tumour that typically occurs in the epidermis. The Skin Cancer Foundation lists five typical characteristics of BCC as:

- Persistent, non-healing open sore that bleeds, oozes or crusts and remains open.
- Reddish patch or irritated area.
- A shiny bump or nodule that may appear pearly or translucent and is often red or white. May also be tan, black or brown.
- Pink, slightly elevated rolled borders with crusted indentation in the centre.
- Scar-like areas that are white, yellow or waxy. Often with poorly defined borders.

Expert opinion affords the further information that these growths typically result in the formation of a "crater" like structure with the surrounding spot slightly elevated and that the skin surface may appear broken.

3.3.2 Squamous cell carcinoma

The second most common form of skin cancer, this growth arises in the squamous cells that are found in the upper layer of the skin. Common characteristics for this type of growth include:

- Scaly red patches with irregular borders.
- A wart-like or open sore appearance that is likely to crust or bleed.
- An elevated growth with a central depression or crater like appearance.

Further to these characteristics, it was noted that these lesions may appear to be relatively large in size and again exhibit a "crater like" appearance with raised surroundings and a central depression.

3.3.3 Seborrheic keratosis

This benign skin growth can be considered very common among people over 40 years of age. Typical features of these lesions are

- Resemble flattened or raised warts.
- Exhibit a variety of colours, from pink or yellow through to brown or black.
- Often described as having a "pasted-on" appearance due to the fact that only the top layers of the epidermis are involved.

3.3.4 Benign melanocytic naevus

More commonly known as birth marks or moles, this class of usually harmless skin lesion is very common and may be present at birth but more usually begin to grow during childhood (although new ones can appear at any age, sometimes in crops). Visually, these lesions tend to be:

- Dark brown or black in colour (due to presence of high melanin concentration).
- Generally circular or oval in shape.
- Typically small in size but with noted exceptions.

Further to the characteristics noted by the Skin Cancer Foundation, it is suggested that Melanocytic nevi may be flat or slightly raised on the surface and portray a relatively smooth surface texture.

3.3.5 Actinic keratosis

As previously noted, this lesion class may be thought of as a pre-cancer and is often the result of prolonged exposure to sunlight. Characteristic symptoms of this class are noted as:

- May be small in size in initial stages, but grows gradually.
- Appears to be crusty, scaly or bumpy on the skin surface.
- Colour may vary and is typically tan, pink red or a combination of these.
- Tendency to become inflamed with surrounding redness or bleeding.

Further characteristics highlighted by a dermatological expert include a reasonably small size and often the lesion will contain a white keratin material.

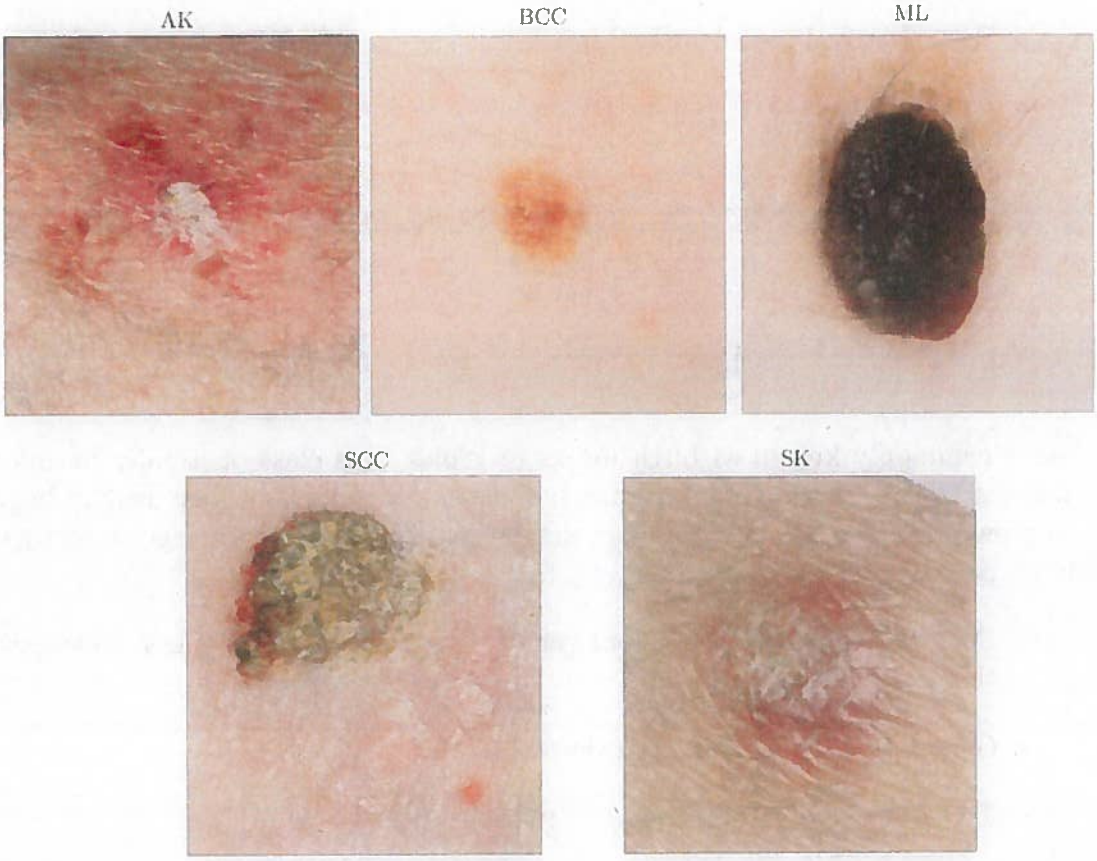


Figure 3.3: Sample images of skin lesions from the provided data set.

3.4 Discussion

The provided data set has a number of noted advantages and disadvantages. Since the data is captured from a real clinical environment, we believe that the examples are highly representative of the various classes of skin cancer both in terms of variation between samples of the same class and ratios with which various classes occur in a clinical setting (with the exception of the ML class which it was thought may be over represented). This will help to ensure our generality of testing. Using a real patient clinic image set also provides an independent and challenging source of test data.

The main drawbacks to this source of data are the relatively low number of total samples and the fact that these samples are not shared examples of the field, making it difficult for us to directly compare our system to related work or have anyone independently reproduce our results. The lack of a definitive set of shared examples in this field was highlighted in Section 2. Also, since the data set is not synthetic we are not able to systematically vary any domain characteristics of interest. Finally we note that not all diagnoses were verified by pathology, so there may be some labelling errors in the training data.

4. System methodology and implementation

The principal goal of the project was to investigate the benefits of surface shape information when classifying images of skin cancer. Towards this goal, a system for the processing, analysis and classification of skin images was developed. This section outlines the main methodologies and processes made use of by our system.

4.1 Preprocessing

4.1.1 Range data preprocessing

It was noted that the lesion samples were captured from a variety of curved body surfaces. On inspection, this often caused the presence of interesting local surface texture to be obscured by global structure. To correct for global surface structure, the range data representing the skin surface was rotated to be approximately perpendicular to the viewer. The technique used to achieve this is described below.

4.1.2 Global orientation

The core idea involves rotating the global surface until it is facing the viewer. We assume the viewer to be directly above the skin surface i.e. in the direction $[0\ 0\ 1]^T$. This involves rotating each pixel as follows:

1. Let each pixel be the vector $p(i,j) = (x(i,j)\ ,\ y(i,j)\ ,\ z(i,j))$
2. Rotate each point as $p'(i,j) = R \cdot p(i,j)$ where R is an appropriate rotation matrix

The rotation matrix R is found for each sample by constructing a 3D basis for the surface consisting of the orthonormal vectors $\{\mathbf{n}, \mathbf{q}, \mathbf{r}\}$ where $\mathbf{n}$ is the global surface normal to an estimated plane fitted to the observed data and $\mathbf{q}$, $\mathbf{r}$ are chosen such that $\mathbf{n} \cdot \mathbf{q} = \mathbf{n} \cdot \mathbf{r} = \mathbf{q} \cdot \mathbf{r} = 0$.

We find $\mathbf{n}$ by estimating the global surface normal to the least squares best-fitting plane through the points on the surface corresponding to *normal skin* (see Section 4.3.1 for *normal skin* definition). This can be regarded as a first order approximation to the surface normal direction at the central point of the surface. This normal vector can be conveniently found by constructing appropriate matrices containing the surface point data and making use of Singular Value Decomposition as follows.

Define matrix M to contain the surface x y z data such that:

$$M = \begin{pmatrix} x_1 & x_2 & \ldots & x_n \\ y_1 & y_2 & \cdots & y_n \\ z_1 & z_2 & \ldots & z_n \\ 1 & 1 & \ldots & 1 \end{pmatrix}$$

Let $S = M \cdot M^T$

Components of the surface normal to the estimated plane can then be derived from the fourth column of the matrix V where:

$$svd(S) = [U, D, V]$$

$$\mathbf{n} = [v_{1,4}\ v_{2,4}\ v_{3,4}]^T$$

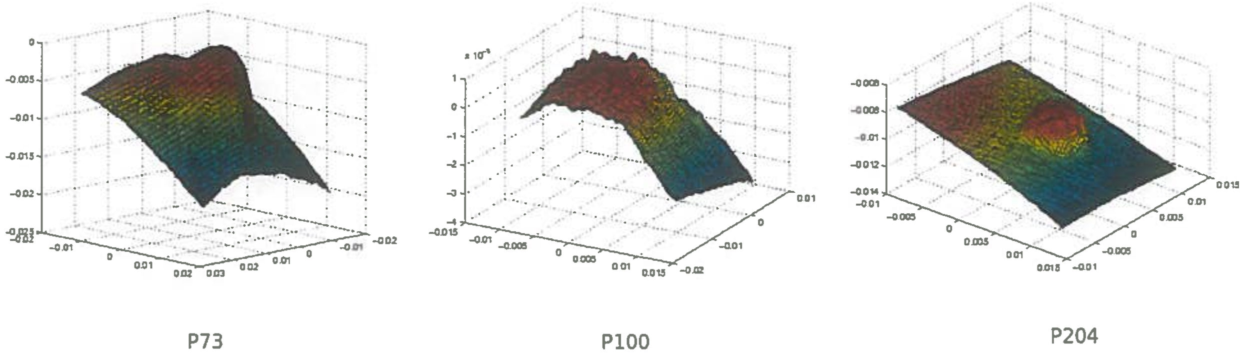
Finally, we are able to solve for the rotation matrix R using:

$$R = [\mathbf{q}\ \mathbf{r}\ \mathbf{n}]^{-1}$$

Two interesting problems of note here were that a matrix R might be found that also implements a *mirror image* or a matrix which resulted in the lesion sample being *inverted* in the z-axis. The first of these problems can be overcome by comparing the signs of a vector triple product (4.1). If the signs are opposite, then $\mathbf{q}$ is replaced with $-\mathbf{q}$ when solving for R. A solution to the invertion problem is to simply check the sign of the z component of the normal vector $\mathbf{n}$ and invert it if it is found to be negative.

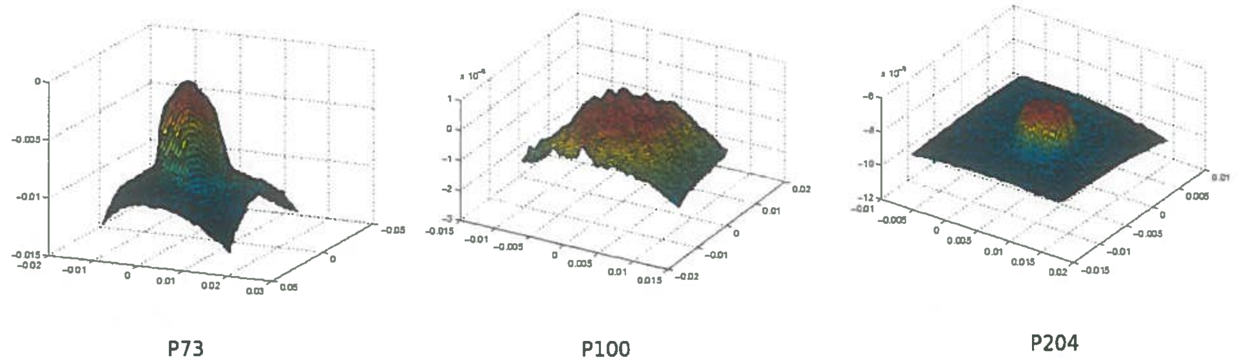
$$sign(([1\ 0\ 0]^T \times [0\ 1\ 0]^T) \cdot [0\ 0\ 1]^T) = sign((R \cdot [1\ 0\ 0]^T \times R \cdot [0\ 1\ 0]^T) \cdot R \cdot [0\ 0\ 1]^T) \tag{4.1}$$

Applying the rotation R to each point results in the corrected depth channel of the data points for each sample. Plots of the data before and after these global rotations are applied are shown in Figures 4.1 and 4.2.



P73 P100 P204

Figure 4.1: Original depth data for three lesion samples



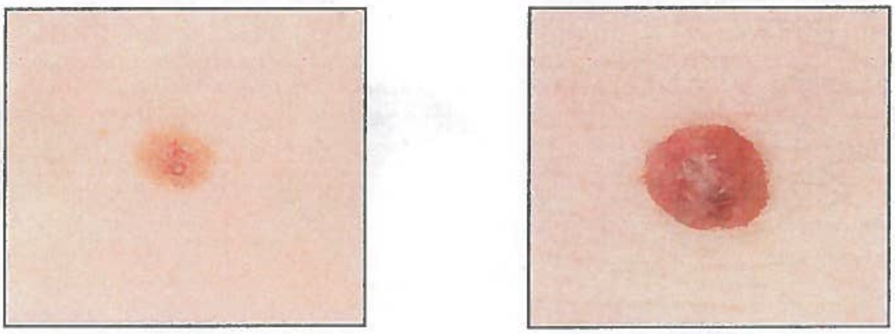
P73 P100 P204

Figure 4.2: Depth data after global rotation transform

4.2 Image sample selection and cropping

In some instances, the data set contained multiple intensity images and depth data samples pertaining to the same skin lesion. In such cases, a method was needed to select the best available sample. To this end, the depth data was visualised in an attempt to select the optimal sample, in terms of the quality of the data captured. The main technique used for visualisation was to derive Cosine Shaded images from the depth data which proved to be useful in assessing the quality of the sample. This technique is detailed in the following section.

Some patient samples were discarded completely at this early stage as upon viewing the Cosine Shaded visualisation it was clear that the 3D image capture process had not performed correctly or had been heavily affected by a profusion of hair follicles which would tend to disrupt the stereo recovery process.



P126

P103

Figure 4.3: Intensity images

4.2.1 Cosine Shading

Cosine Shading is a simple shading algorithm commonly used in visualisation software. The algorithm utilises the cosine of the angle θ between the surface normal vector $\mathbf{n}_s$ at each point in the image s and a normalised light source vector $\mathbf{I}$ in order to calculate an estimation of the brightness at this point on the surface. Viewing the skin surfaces as a set of brightness values in this form proved helpful in illustrating the underlying structure present in the depth data. The cosine shaded images produced in this project assumed that the light source was positioned directly above the image therefore $\mathbf{I} = [0\ 0\ 1]^T$ and the surface normal to each point was calculated using a $N \times N$ window surrounding the point using the least squares plane estimation method described previously in Section 4.1.1. Experimentally, values of $N = 3$ were found to produce the most descriptive images. The brightness value at each pixel I_s is then calculated as in Equation (4.2). Examples of colour lesion image samples and cosine shaded images produced from their corresponding range data are shown in Figures 4.3 and 4.4.

$$I_s = \mathbf{I} \cdot \mathbf{n}_s = |\mathbf{I}||\mathbf{n}_s|\cos\theta \tag{4.2}$$

Figure 4.3 displays two examples of skin samples from the Mole class. It can be noted in Figure 4.4 that the Cosine Shaded image produced by sample P126 can be seen to be fairly flat and uniform. This would suggest that there is not a lot of descriptive information available from the depth data for this sample. Sample P103 on the other hand exhibits noticeable three dimensional structure in the Cosine Shaded image and is an example of a good specimen with respect to the collected range data. The presence or lack of descriptive surface structure is something that would otherwise prove difficult to deduce from the intensity

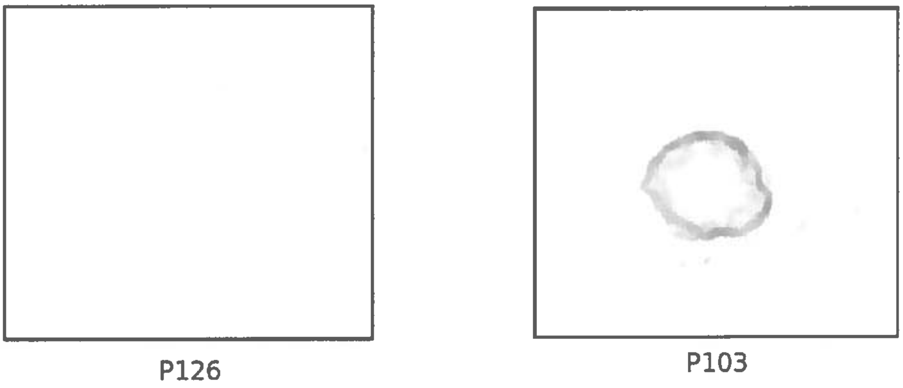


Figure 4.4: Corresponding Cosine Shaded images

images alone.

4.2.2 Resolution cropping

As previously mentioned the original intensity images were captured at a resolution of 3456×2304. In a pre-emptive effort to diminish the computational expense of extracting features from high resolution image data, both the intensity and depth images were cropped at this stage to a size of typically ~700×700 pixels, retaining the main area of interest (the skin lesion) and an area of surrounding skin.

4.3 Image segmentation

Segmentation refers to the process of extracting meaningful regions out from the image background. The initial preprocessing step for the data in this project involves segmenting the lesion in each image from the background (skin) to aid in the following feature extraction steps. Accurate segmentation of the regions of interest can be considered a crucial step in the skin spot classification process.

Traditionally, automated segmentation techniques can be categorised into two classes: those that employ region-finding algorithms versus those that employ contour-detection algorithms. Most classical region-finding algorithms involve partitioning a grey level histogram in such a way that appropriate thresholds for segmentation can be easily found. However, studies indicate that for complex images such as those of skin tissue structure, simple thresholding techniques based on a globally determined values may not work well [15].

Since the focus of this project was on defining features and classification, we chose to perform the segmentation task manually with a free hand tool which was created for the purposes of the project.

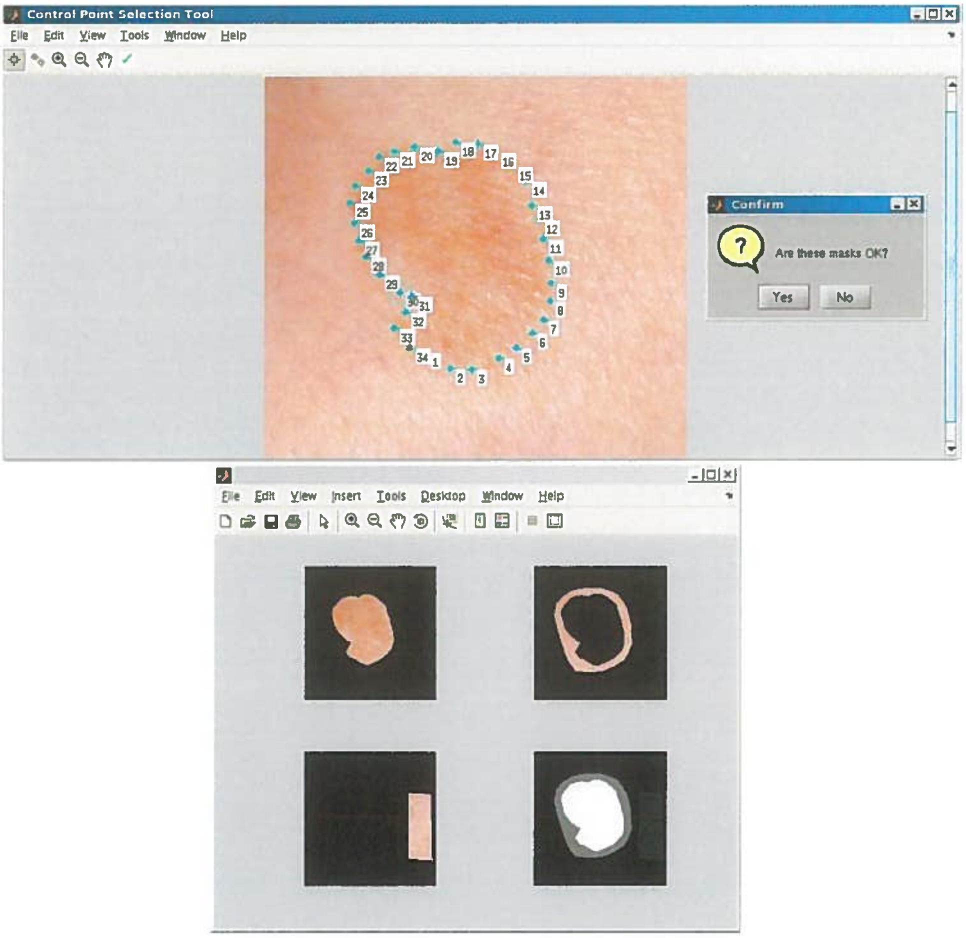
Manual segmentation may be thought of as a highly reliable segmentation method, especially when performed by a dermatologist or similar domain expert. The caveat here is that segmenting a large collection of images by hand can be a time consuming (and rather laborious!) process. It is noted that successful solutions to the problem of automated segmentation also exist in this and related domains. For example Ganster et al report favourable results using adaptive local thresholding methods in [19] for the purpose of segmenting ELM based dermatological images.

4.3.1 Image segmentation tool

An image segmentation tool was designed and implemented for the purpose of manual segmentation. The tool loads each sample intensity image and displays it. The user is then able to select a set of (x, y) control points on the image which are used to define the perimeter of the lesion and other areas of interest. The polygon defined by connecting these control points is then used to produce three binary masks corresponding to areas of the image containing the lesion, a patch of "normal skin" and a small region surrounding the lesion defined as "uncertain" which typically represents an ambiguous zone found at the normal skin / lesion border. This divides the point/pixel data into three sets. If an automated segmentation technique were being used, selection can be additionally performed by a specified amount of *erosion* (using standard morphological operators), the effects of which are to eliminate any ambiguous "uncertain" points. It is with our hand cropped "uncertain" mask that we attempt to simulate that effect in this project. These binary masks can then be used to selectively extract information from the defined areas of interest in both the intensity and depth data. The segmentation tool and sample output can be found in Figures 4.5 and 4.6.

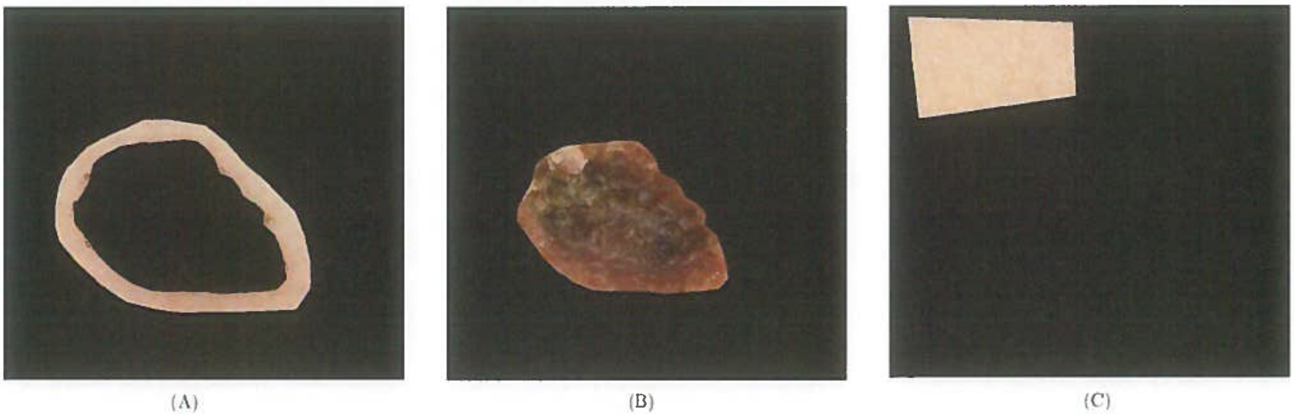
4.4 Summary

The previous series of sequential operations and preprocessing techniques are carried out in an attempt to aid us when extracting meaningful patterns and features that may be present in the data. After performing the global surface orientation, cropping and segmenting the intensity data, we are in a position where we are able to extract potentially useful properties with which we can describe the various classes of skin lesion.


Control Point Selection Tool
File Edit View Tools Window Help
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34
Confirm
Yes
No
File Edit View Insert Tools Desktop Window Help

Are these masks OK?

Figure 4.5: Image segmentation tool


(A)

(B)

(C)

Figure 4.6: Examples of the "uncertain" (A), skin lesion (B) and "normal skin" (C) masked areas of a sample lesion

5. Colour and depth feature extraction

The project used a number of different parameters from the intensity and depth data in the Bayesian classifier described in chapter six. To some extent, these features try to reflect the parameters used in medical diagnosis. Some example feature categories are *size* (area, perimeter, polar measures and bounding rectangle) and *shape* (formfactor, roundness, compactness) [32] and *colour* (minimum, maximum, average and variance of image RGB channels). The novel features we calculate are based on the image depth data. Example *depth* features include spot height, average spot roughness and peak and pit densities. These features are then used as input to machine learning techniques for the purpose of image classification. Here we formally define the features implemented, describe how they were extracted from the data and examine their variance and class distributions within the available data set.

5.1 Feature: Mean spot height

One of the initial features derived from the depth data is what we define as *Mean spot height*. This feature was defined in an attempt to represent the amount that a skin lesion protrudes from the surface in relation to the surrounding skin. It was predicted that a feature representing the average height of a skin lesion would prove to be descriptive given that a number of physical lesion properties (cf. Section 3.3) are height related (e.g. the noted "elevated growth" and "crater like" characteristics of SCC).

5.1.1 Feature definition

Calculating this feature involves taking the mean z value of depth data points residing inside the set of points S belonging to the *skin lesion*, as defined by the lesion binary mask and finding the difference with respect to the mean z value of the set of points T from the surrounding *skin* area, as defined by the "normal skin" binary mask (cf. Section 4.3). This measure is defined in Equation (5.1).

$$\Delta_{\text{mean spot height}} = \frac{1}{|\mathrm{S}|} \sum_{z_i \in S} z_i - \frac{1}{|\mathrm{T}|} \sum_{z_i \in T} z_i \tag{5.1}$$

5.1.2 Algorithm verification

To verify the correctness of the implemented calculation, synthetic data was created to simulate depth data captured from a lesion surface. The lesion area was simulated by a hemisphere of radius $r = 0.1$ and the normal skin area was represented by a flat plane with a small amount of Gaussian noise added ($\mu = 0$, $\sigma = 0.001$) to simulate small imperfections typically found on the epidermis. A plot of the synthetic data can be seen in Figure 5.1. Given that the dimensions of the hemisphere structure were known, it was possible to predict the values that the mean spot height extraction algorithm would produce and compare these with the observed values. For example, the mean lesion height can be predicted using Equation (5.2). The predicted and experimental results produced by this feature using the synthetic data can be found in results Table 5.1 below. It is noted that the experimental value is slightly below the predicted value. This may be explained by the lesion mask perimeter being slightly larger than that of the synthetic hemisphere, therefore a small amount of relatively flat skin at the border of the spot is taken into account when calculating the average height, slightly lowering the observed value.

$$\text{Mean lesion height} = \frac{\frac{2}{3}\pi r^3}{\pi r^2} = \frac{2}{3}r \tag{5.2}$$

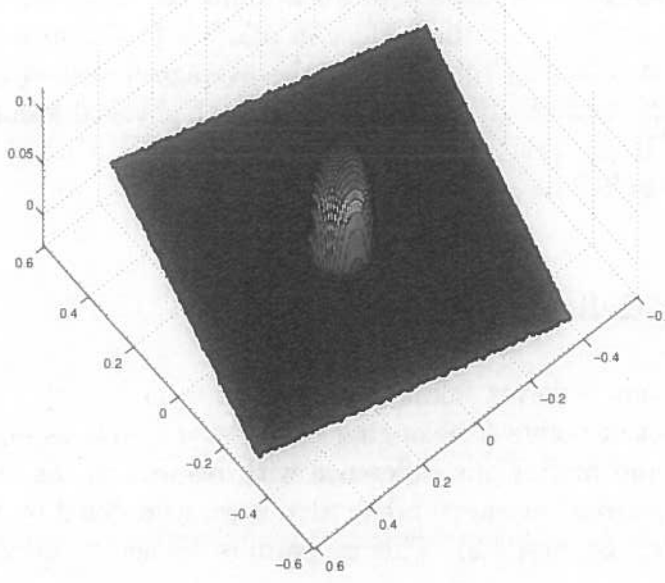


Figure 5.1: Synthetic skin lesion depth data

	Mean lesion height	Mean skin height	Δmean spot height
Predicted	0.0666	ϵ	0.0666 - ϵ
Observed	0.0576	2.3313×10^{-7}	0.05759

Table 5.1: Synthetic data mean spot height values

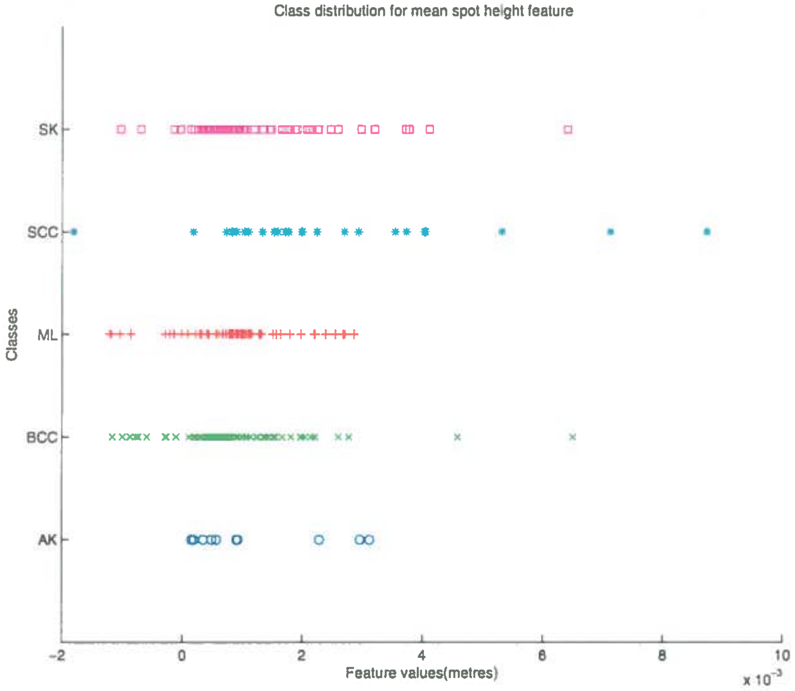
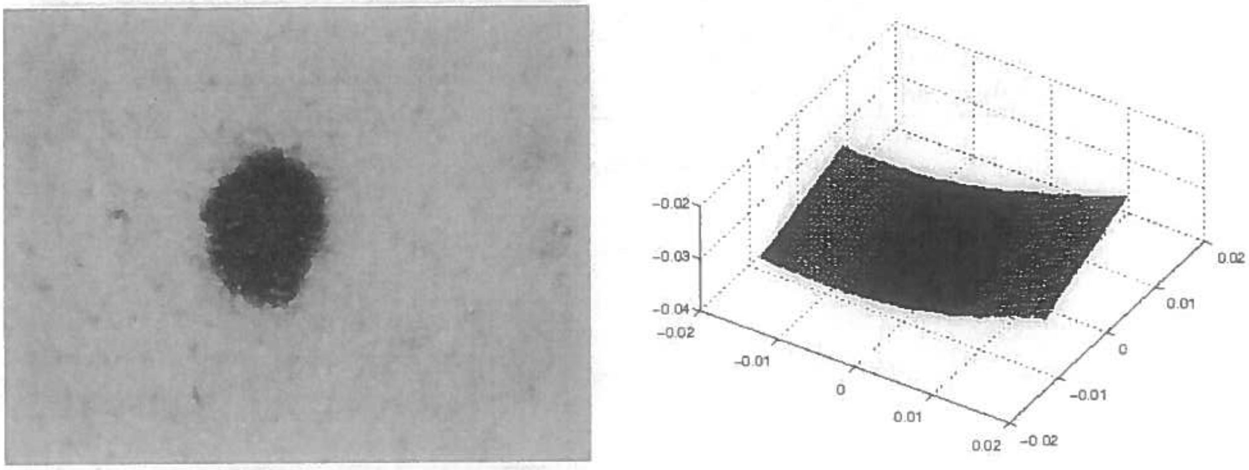


Figure 5.2: Class distributions for mean spot height feature

5.1.3 Observations

The class distributions of the data set in the feature space defined by this property are plotted in Figure 5.2. The classes do appear to exhibit some degree of separability in this space. Classes SCC and SK have fairly long right hand tails, exhibiting a large number of samples with values greater than $2mm$ in contrast to the other three classes. This indicates that SK and SCC samples may tend to physically protrude from the skin more, making this feature a useful descriptor for these classes in particular.

It can also be noted from this plot that a number of lesions from the classes BCC and ML have a negative mean spot height, indicating an overall depressed or concave global lesion structure. The data point belonging to the ML sample on the furthest left of the distribution (largest negative mean spot height) was investigated and the corresponding intensity and depth data can be seen in Figure



P177

Figure 5.3: Intensity and range data for ML sample with negative feature value

5.3. The intensity image gives the impression that this mole is fairly small and flush with the skin, providing little in the way of useful local depth information (this was confirmed by inspecting the corresponding cosine shaded image). The depth data shows how the global structure and curvature of the body part has eclipsed any local subtleties, (even after global orientation correction) accounting for this uncharacteristic feature value.

The class distributions appear to be fairly unimodal with the exception of the class AK where a lack of available samples makes an accurate prediction about the shape of the distribution difficult. BCC also appears to exhibit a slightly taller and narrower peaked distribution than the others.

5.2 Feature: Roughness ratio

A second family of metrics making use of the depth data and also the colour intensity channels is defined as *Roughness ratio*. Here we hope to express any macro variation of the lesion surface area in the *z-depth* channel and *red, green, blue* colour channels. This family of features attempts to take advantage of the fact that the skin classes studied typically exhibit varying levels of surface roughness. For example, in [3] it is noted that lesions belonging to the classes SCC and BCC are typically "crusty" with irregular borders while benign moles (represented by the class ML here) are described as having a relatively smooth surface.

5.2.1 Feature definition

This feature is derived for each of the four available channels of information by taking the standard deviation σ of the set of points S in the *lesion area* of each sample and normalising by the standard deviation of the set of points T in each sample consisting of *normal skin*. For smoother lesions, it is expected that this ratio will be close to 1, i.e. there is not much difference between the roughness of the lesion area and surrounding normal skin.

This measure is defined in Equation (5.3) where σ_X^c is the first standard deviation of the points in area X from information channel $c \in \{\textit{red,green,blue,z-depth}\}$.

$$\text{Roughness ratio}_c = \frac{\sigma_S^c}{\sigma_T^c} \tag{5.3}$$

5.2.2 Algorithm verification

The correctness of the algorithm implemented to calculate roughness values was again verified by the use of synthetic data. For the intensity channels (*red, green, blue*) a synthetic image was created with pure red (RGB $[255, 0, 0]$) and green (RGB $[0, 255, 0]$) areas. Discrete uniformly distributed values $(0 - 255)$ were added to each pixel in the *red, green, blue* channels to simulate intensity roughness and variation. A similar synthetic image was created for the *z-depth* channel again using uniformly distributed values $(0 - 100)$ to simulate variation in the *spot* area and *skin* area. An example of the synthetic colour data is displayed in Figure 5.4.

The standard deviation of a discrete uniform distribution from values a to b is defined in Equation (5.4) and can be used to calculate expected values for the synthetic data with respect to this feature family. Predicted and observed results can be seen in Table 5.2. The difference between the theoretical and observed values was found to be small in all cases (≤ 0.052). The infinity value is the green roughness ratio entry is explained by the fact that all pixels in the skin area will have the maximum value (255) in the green channel, giving a standard deviation (and therefore divide by) zero.

$$\sigma = \sqrt{\frac{(b-a)^2}{12}} \tag{5.4}$$

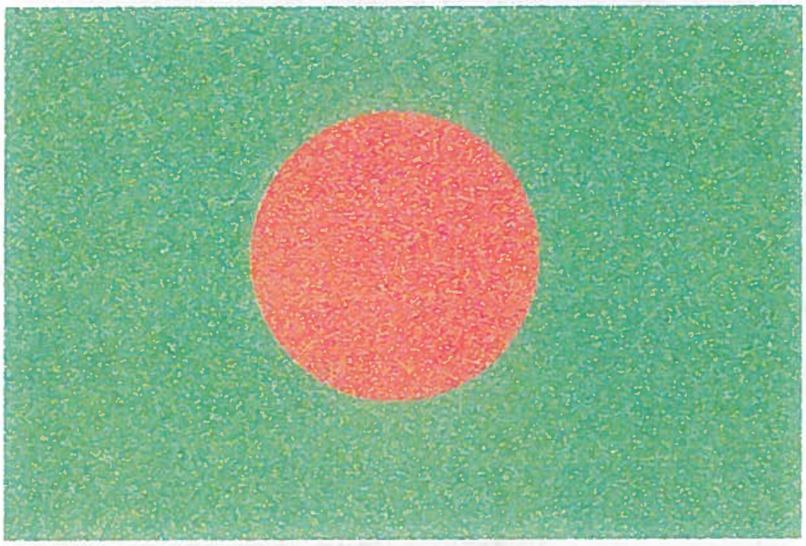


Figure 5.4: Synthetic intensity data with uniform *red green blue* channel noise

Roughness ratio				
Roughness ratio	$\frac{\sigma_S^r}{\sigma_T^r}$	$\frac{\sigma_S^g}{\sigma_T^g}$	$\frac{\sigma_S^b}{\sigma_T^b}$	$\frac{\sigma_S^z}{\sigma_T^z}$
Predicted	0.0	∞	1.0	1.0
Observed	0.0520	∞	1.0041	0.9996

Table 5.2: Spot area standard deviations

5.2.3 Observations

The distributions for this roughness feature in the *z-depth* channel are plotted in Figure 5.5. Corresponding plots for the *red, green* and *blue* channels can be found in the Appendix. It can be seen from the distributions in Figure 5.5 that there is significant overlap amongst the five classes, with each class typically displaying a heavy concentration of samples near a ratio of 1.0, indicating the existence of globally smooth samples in all classes.

A number of outliers displaying high roughness ratio values can be seen in the plot but this does not appear to be class specific. The shape of the distributions would suggest that this feature in the *z-depth* channel (and indeed the three colour channels) do not exhibit sufficient variations across the skin classes to aid reliable classification.

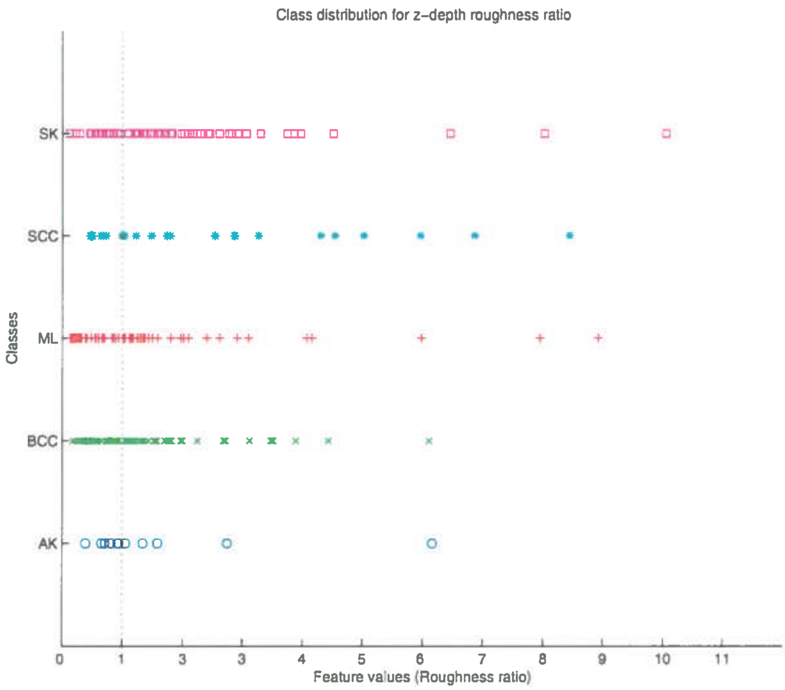


Figure 5.5: Class distributions for Roughness ratio in *z-depth* channel

5.3 Feature: Mean intensity ratios

A set of nine further features were derived from the mean values of the intensity channels (*red, green, blue*). This family of features was implemented in an attempt to automatically compute quantities which emulate the "colour" aspect of the ABCD clinical diagnosis rule criteria discussed previously. Variegated colour based features are cited in [25] as one of the most predictive class of features in the classification of malignant melanoma lesions. With this feature family, we explore whether this success can be extended to the classes of skin lesion studied here.

5.3.1 Feature definition

These features involve calculating mean values in each of the *red, green* and *blue* colour channels. Mean values are computed for each intensity image in areas defined by both the spot and skin binary masks, in a similar fashion to the previous roughness properties. Normalised spot-to-skin ratios are then used to define the features. The nine ratios used in this feature set are listed in

Equation (5.5) where μ_{c_1S} , μ_{c_2T} are the mean intensity values of channel c_1,c_2 $\in \{red, green, blue\}$ in the *spot* and *normal skin* areas of the image, respectively.

$$\frac{\mu_{rS}}{\mu_{rT}}, \frac{\mu_{rS}}{\mu_{gT}}, \frac{\mu_{rS}}{\mu_{bT}}, \frac{\mu_{gS}}{\mu_{rT}}, \frac{\mu_{gS}}{\mu_{gT}}, \frac{\mu_{gS}}{\mu_{bT}}, \frac{\mu_{bS}}{\mu_{rT}}, \frac{\mu_{bS}}{\mu_{gT}}, \frac{\mu_{bS}}{\mu_{bT}} \tag{5.5}$$

5.3.2 Algorithm verification

The correctness of the extraction method for this feature set was verified by making use of similar synthetic colour data to that found in Section 5.2.2. A synthetic colour image similar to that of Figure 5.4 was hand segmented using our spot segmentation tool under the assumption that the circular pure red area of the image represented the lesion and the pure green area represented the normal skin area. From the resulting binary masks the ratio values for this feature set could be predicted[1] and compared to the observed values. Predicted and experimental results are found in Tables 5.3 and 5.4. The small discrepancies are explained by the fact that a few colour pixels on the skin / spot border were likely partitioned into the the wrong mask due to the imperfect nature of hand segmentation.

		Spot		
		$R = 255$	$G = 0$	$B = 127.5$
Skin	$R = 0$	∞	∞	∞
	$G = 255$	1.0	0.0	0.5
	$B = 127.5$	2.0	0.0	1.0

Table 5.3: Predicted $\frac{Spot}{Skin}$ intensity ratios

		Spot		
		$R = 255$	$G = 0$	$B = 127.5$
Skin	$R = 0$	∞	∞	∞
	$G = 255$	0.9977	0.0023	0.5010
	$B = 127.5$	1.9924	0.0046	1.0005

Table 5.4: Observed $\frac{Spot}{Skin}$ intensity ratios

5.3.3 Observations

The class distribution for one of the most promising features ($\frac{\mu_rS}{\mu_gT}$) from this family is plotted in Figure 5.6. Plots of the further eight features from this set

[1] Predicted blue channel "noise" values are the expectation of a uniformly distributed variable (0-255) $E(B) = \frac{255}{2}$

can be found in the Appendix. This ratio feature and indeed others in this family do seem to exhibit some separation between the studied classes. In particular the classes SK and ML have a significant number of samples found below the data set mean ($\mu = 1.1309$), with these distributions also displaying distinctive long left hand tails.

The remaining three classes SCC, BCC and AK all exhibit noticeably higher distribution means with AK in particular (all samples bar one) lying completely above the mean value for the data set, which would suggest high red channel values for this class. This agrees with typical Actinic Keratosis visual characteristics which were noted to include inflammation with surrounding redness or bleeding. These characteristics were confirmed to be consistant with the available samples by visual inspection and it is therefore thought that this feature and others from this family may prove useful for lesion identification.

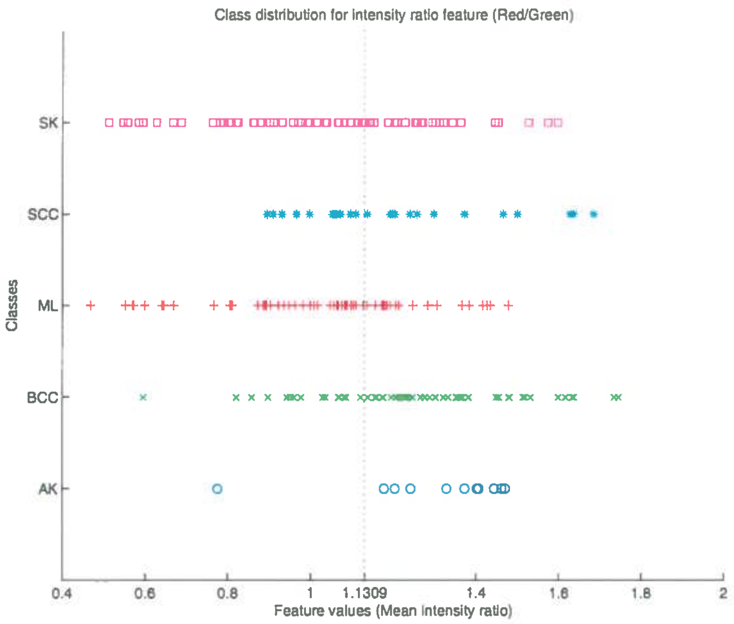


Figure 5.6: Class distributions for Mean intensity ratio $\frac{\mu_r S}{\mu_g T}$

5.4 Feature: Peak and pit densities

Textural features can be loosely divided into two categories: structural and statistical [27]. Structural features try to determine the primitives with which the texture is composed while statistical features typically compute local features at each point in the image and derive a set of statistics from the distributions of these local features.

Statistical methods can be further classified as first-order (one pixel), second-order (two pixel) and higher-order (three or more pixel) statistics. First-order statistics ignore the spatial interaction between image pixels while second and higher-order statistics estimate properties occurring at specific locations relative to each other.

Here we define a higher-order statistical feature set which counts the number of *peak* and *pit* values in each information channel over the region defined by the spot binary mask. With these features we attempt to capture descriptive local texture information from the lesion samples. *Peak* and *pit* values are formally defined in the following section.

Since the coarseness of an image is not absolute but depends on the scale at which the image is processed or viewed, several levels of granularity were experimented with when extracting this family of features.

5.4.1 Feature definition

A *peak* (or *pit*) value is defined to be a data point that has no surrounding points in a neighbourhood of cardinality N with a greater (or lower) value. A qualifying peak value $x_{2,2}$ and corresponding neighbourhood region for $N = 9$ can be seen in Equation (5.6).

$$x_{1,1}, x_{1,2}, x_{1,3}, x_{2,1}, x_{2,3}, x_{3,1}, x_{3,2}, x_{3,3} < x_{2,2} \tag{5.6}$$

$x_{1,1}$	$x_{1,2}$	$x_{1,3}$
$x_{2,1}$	$x_{2,2}$	$x_{2,3}$
$x_{3,1}$	$x_{3,2}$	$x_{3,3}$

A statistical feature is then defined for each information channel $c \in \{red, green, blue, z\!-\!depth\}$ as follows:

$$\text{Peak and pit density}_c = \frac{\#\text{peaks}_c + \#\text{pits}_c}{\text{spot area}} \tag{5.7}$$

The image data was passed through a Gaussian lowpass filter with varying standard deviations ($\sigma = 0.5, 1, 2$) and the feature set extracted at each of the resulting levels of granularity.

5.4.2 Algorithm verification

To verify the correctness of these features, synthetic intensity and range data was engineered to contain representative local texture information. Depth data was created using a linear combination of a cosine and sine waves to produce a "bumpy" surface and then binary masked as a lesion. Synthetic intensity data was created by hand, by adding small *red, green* and *blue* areas to a constant brown background. Peak and pit areas could then be counted by visual inspection and compared to the results obtained by the feature extraction algorithm. Sample test range data is found in Figure 5.7 and the corresponding intensity image can be seen in Appendix. Test results can be seen in Tables 5.5 and 5.6. It can be noted that no blue pits were found in the intensity test data due to a brown background colour (RGB: $128, 64, 0$) being used meaning no points had a strictly lower blue channel value.

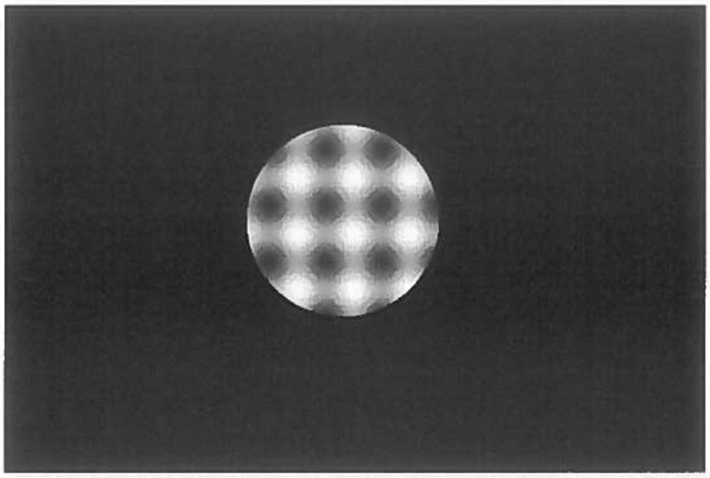


Figure 5.7: Synthetic textured depth data

	Red peaks	Green peaks	Blue peaks	z-depth peaks
Predicted	10	10	10	11
Observed	10	10	10	11

Table 5.5: Synthetic data peak counts

	Red pits	Green pits	Blue pits	z-depth pits
Predicted	20	20	0	8
Observed	20	20	0	8

Table 5.6: Synthetic data pit counts

5.4.3 Observations

The plot in Figure 5.8 show the distributions in the depth channel with different levels of the Gaussian filter applied. A large amount of overlap among the classes is exhibited by both data plots in this information channel. The distributions are very broadly spread with nearly every class containing samples across the spectrum of obtained values. This trend of high overlap is consistant across all four information channels and in all explored levels of Gaussian filtering. Further work might take the direction of exploring a wider range of Gaussian σ values, as the effect of the range experimented with here can be seen to be minimal.

It was predicted that classes of typically smoother lesions such as moles (ML) would produce a distinctive distribution under such texture properties, however this lack of class separability seems to make the features unlikely to contribute significantly to accurate classification.

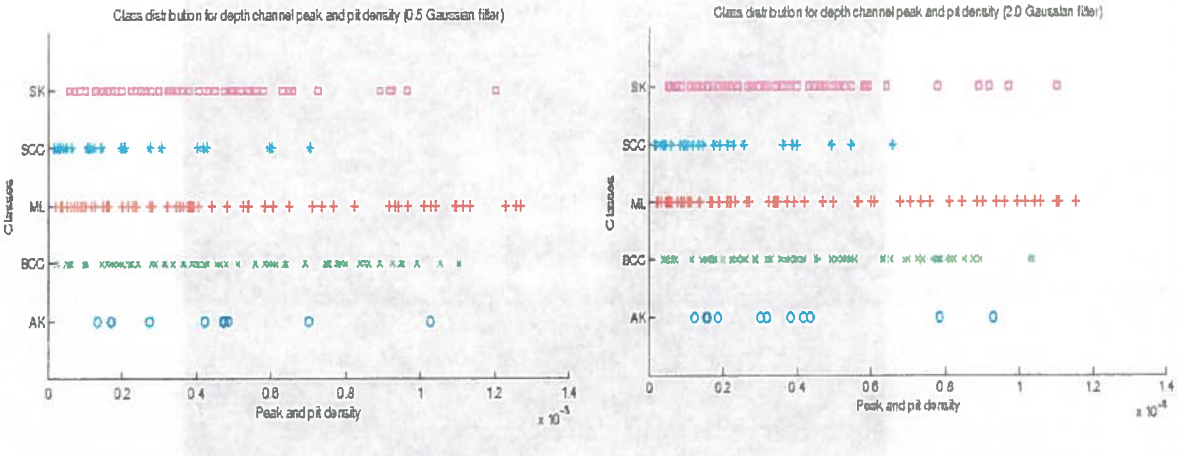


Figure 5.8: Class distributions for depth channel peak and pit density (Gaussian image filter Left: $\sigma = 0.5$ Right: $\sigma = 2.0$)

5.5 Feature: 3D moment invariants

Moment based features are used in many computer vision applications and can be thought of as weighted averages of the image pixel intensities (or functions of those moments, chosen to have some useful interpretation). Originally defined in two dimensions by Hu [21], the technique can be extended to compute moment invariants of 3D point distributions that are invariant to translation, rotation and scale thus providing a measure of the spatial distribution of mass in a 3D shape. Here we make use of the available depth based data to calculate modified versions of three second order moment invariants originally due to Sadjadi and Hall [33].

5.5.1 Feature definition

Let $(\bar{x}, \bar{y}, \bar{z})$ be the centroid of the object.

Let S be the set of points defined by the binary spot mask.

Let

$$\rho(x, y, z) = \begin{cases} 1 & \text{if } (x, y, z) \in S \\ 0 & \text{if } (x, y, z) \notin S \end{cases}$$

The 3D moments of order $n = p+q+r, n \in N$ of the 3D density function $\rho(x, y, z)$ are defined by

$$\mu_{pqr} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x - \bar{x})^p(y - \bar{y})^q(z - \bar{z})^r \rho(x, y, z) dxdydz \tag{5.8}$$

The continuous moment definition found in Equation (5.8) can be approximated to our digital depth data by using summation in place of integration. From this we define three second order moment features that are invariant to translation and rotation:

$$\begin{aligned} I_1 &= \mu_{200} + \mu_{020} + \mu_{002} && (5.9) \\ I_2 &= \mu_{200}\mu_{020} + \mu_{200}\mu_{002} + \mu_{020}\mu_{002} - \mu_{110}^2 - \mu_{101}^2 - \mu_{011}^2 && (5.10) \\ I_3 &= \mu_{200}\mu_{020}\mu_{002} + 2\mu_{110}\mu_{101}\mu_{011} - \mu_{002}\mu_{110}^2 - \mu_{020}\mu_{101}^2 - \mu_{200}\mu_{011}^2 && (5.11) \\ & && (5.12) \end{aligned}$$

To make Sadjadi and Hall's original moments also invariant to scale we further normalise each moment by a function of the moment order and shape volume:

$$\mu_{pqr} \leftarrow \frac{\mu_{pqr}}{(Volume)^{(p+q+r+3)/3}} \tag{5.13}$$

Finally we define the three moment features as monotonic functions with a scaling constant C:

$$\begin{aligned} J_1 &= \log_{10}(I_1) + C_1 && (5.14) \\ J_2 &= \log_{10}(I_2) + C_2 && (5.15) \\ J_3 &= \log_{10}(I_3) + C_3 && (5.16) \\ & && (5.17) \end{aligned}$$

5.5.2 Algorithm verification

Verification was carried out by again producing synthetic 3D data to test the moment value calculations. A simple cube of edge length two was used in this case to simulate the lesion surface area. The simple shape and dimensions of the cube aided in the ease with which moment values could be calculated. An example plot of the synthetic data and verification results are found in Figure 5.9 and Table 5.7 respectively. The predicted value for I_1 for the synthetic data is derived here explicitly with I_2 and I_3 calculated analogously.

$$\begin{aligned}\mu_{200} &= \int_0^2 \int_0^2 \int_0^2 x^2 y^0 z^0 dxdydz \\ &= \frac{x^3}{3}\Big|_0^2 \\ &= \frac{8}{3}\end{aligned}$$

Applying scale normalisation: $\mu_{200} \leftarrow \frac{\mu_{200}}{8^{5/3}}$

$$\begin{aligned}\mu_{200} &= \frac{8}{3 \cdot 32} = \mu_{020} = \mu_{002} \\ \therefore\ I_1 &= \frac{8}{96} + \frac{8}{96} + \frac{8}{96} = \frac{1}{4}\end{aligned}$$

	I_1	I_2	I_3
Predicted	0.25	0.0208	0.000625
Observed	0.2484	0.0206	0.000568

Table 5.7: Predicted and observed moment invariant values

5.5.3 Observations

Figure 5.10 shows the class distributions of values calculated for the moment feature J_2 defined in Equation (5.15). The distributions appear to be similar across the considered classes, with the feature values exhibiting a lot of overlap and similar overall distribution shape. Similar results were found for the distributions of J_1 and J_3 (which can be found in the Appendix).

It is noted that the hand segmented spot mask is likely to heavily influence the values that this feature family produces, as the mask essentially dictating the

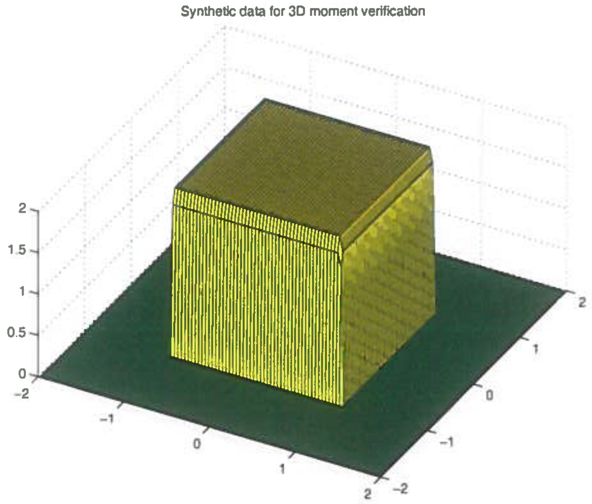


Figure 5.9: Synthetic depth data for 3D moment invariants

shape of the spot in the x and y dimensions. During hand segmentation, the spot mask shape was found to be fairly uniform across all samples processed and therefore does not afford much information about which class a lesion belongs to. It is possible that extracting these features in conjunction with an automated segmentation technique might be able to produce more class distinctive spot mask shapes and in turn, more descriptive invariant moment features.

5.6 Data rescaling

All the features presented in this chapter have different physical units (e.g. average spot height in *meters*) or are without any unit (e.g. *3D moment invariant values*) and, therefore, have different value ranges. To account for this, an objective scaling of the features is achieved by calculating z-scores [19] which are defined in Equation (5.18) where x_{ij} represents the ith sample measure of feature j and μ_j the mean value of all samples for feature j and σ_j is the standard deviation of the samples for feature j.

The z-scores have the statistical properties $\mu_z = 0$ and $\sigma_z = 1$ and are objectively derived without any knowledge about individual features. This dimensionless quantity indicates how many standard deviations an observation is above or below the mean. Converting the features to this form allows us to easily compare different feature distributions (as are discussed in Section 6.2). Standardising the feature values would also allow us to make use of distance based classifiers in future work.

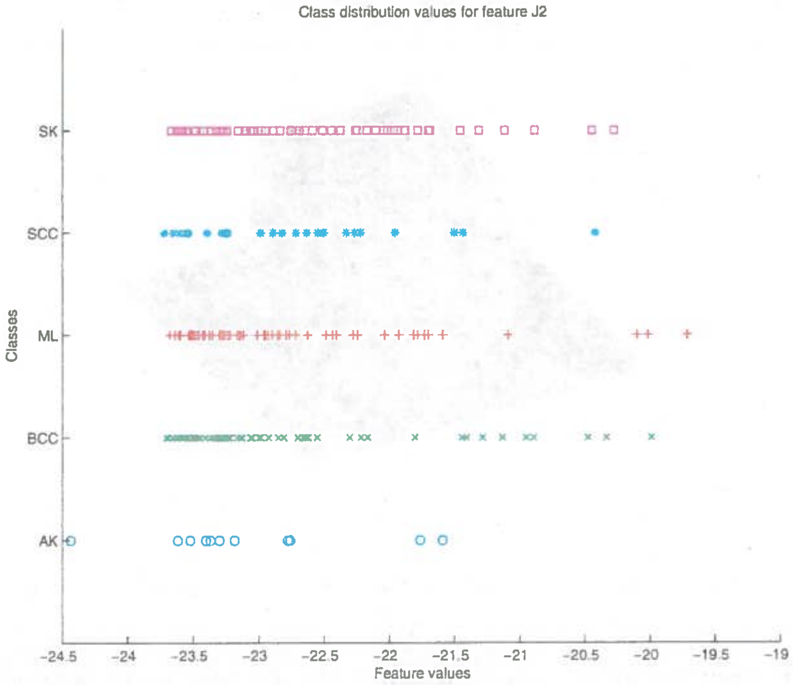


Figure 5.10: Class distributions for feature J_2

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j} \tag{5.18}$$

5.7 Feature selection

We are exploring a real-world classification problem where the underlying class probabilities and class-conditional probabilities are not exactly known. In real-world situations such as this, relevant features are often unknown *apriori* so in this chapter we have introduced many candidate features to represent the skin lesion domain. Unfortunately, as hinted at by some of the observed class distributions in this chapter, many of these features are either partially or completely irrelevant or redundant to the target concept, where irrelevant features do not affect the target concept in any way and redundant features do not add anything new to the target concept [22] (a relevant feature is neither irrelevant nor redundant).

Eliminating these unwanted features has several advantages. If a feature space with a large dimension is used, the performance of the classifier may decrease with respect to execution time and to recognition rate [4]. The execution time increases with the number of features because of the measurement cost. The recognition rate can decrease because of redundant features and of the fact that using fewer features can alleviate the *curse of dimensionality* where a limited number of training samples leads to overtraining.

The second issue especially can be seen to apply to this project, with a noted lack of samples available in certain classes. For example, only eleven usable samples of Actinic Keratosis are available in the data set in the most extreme case. Due to the nature of the probabilistic classifier used (detailed in Chapter 6), we are constrained by the fact that having N training samples per class lets us use a maximum of $N-1$ features for the purpose of classification. It can also be argued that a classifier built with a smaller number of features is preferable to an equally accurate classifier built with the complete set of features.

Due to the fact that some of the defined features are likely not to be informative for classification and that the number of training samples available for building the classifier is substantially smaller than the number of possible features in some classes, a *Feature selection* method is presented here that attempts to pick a relevant and descriptive subset of features for the purpose of classification.

5.7.1 Criterion function

Feature selection methods have been looked at from various angles with [12] noting four categories of Feature selection algorithm.

1. *Idealised*: find the minimally sized feature subset that is necessary and sufficient to the target concept [23].
2. *Classical*: select a subset of M features from a set of N features, $M < N$, such that the value of a criterion function is optimised over all subsets of size M [28].
3. *Improving Prediction accuracy*: the aim of feature selection is to choose a subset of features for improving prediction accuracy or decreasing the size of the structure without significantly decreasing prediction accuracy of the classifier built using only the selected features [24].
4. *Approximating original class distribution*: the goal of feature selection is to select a small subset such that the resulting class distribution, given only the values for the selected features, is as close as possible to the original class distribution given all feature values [24].

Here we present an approach based on the second and third of these categories aiming to find feature subsets that improve the prediction accuracy of the classifier and minimise the cost based metric described in Section 6.1.4. Approaches using this type of evaluation function can be thought of as "wrapper methods". (i.e., the classifier is the evaluation function). As the features are selected using the classifier that later on uses these selected features in predicting the class labels of unseen instances, the accuracy level is expected to be high.

5.7.2 Exhaustive selection

Ideally, a feature selection algorithm would search through all subsets of N features and try to find the optimal one among the competing 2^N candidate subsets according to the evaluation function. This procedure is exhaustive in that it would evaluate every possible combination, trying to find the best one. For the purposes of this project, an exhaustive approach proved too costly[2] and practically prohibitive with even a medium sized original feature set of $N = 30$. With this in mind, a second method based on a greedy heuristic was used in an attempt to reduce computational complexity.

5.7.3 Greedy selection

Due to the high dimensionality of the feature space, a greedy feature selection algorithm was proposed. The algorithm devised for feature subset selection is a version of *Sequential forward selection* (SFS) [13] in that it starts with an empty subset and iteratively adds one feature at a time so as to maximise the intermediate criterion value until the required dimensionality is achieved. A sketch of the method is given in Algorithm 1

The main caveat here is a performance compromise. As with any greedy heuristic making locally optimal decisions, we are not guaranteed to find the global maxima (overall best feature subset). The other drawback of such a simple approach is that it may suffer from so-called nesting of feature subsets which may deteriorate the performance. Attempts to overcome this problem have been made with more recent approaches such as *Sequential floating search* methods [29].

5.8 Summary

In this chapter we have presented the techniques used to extract various features from both the intensity and depth data channels and a selection method to reduce the dimensionality of the feature space for the purposes of classification.

[2] For this project the evaluation of a single feature subset took $\approx$ 15 seconds. With 30 features, an exhaustive search of the space was estimated to take $2^{30} \times 15$ sec $\approx$ 186,500 days!

Algorithm 1 Greedy feature subset selection

```
(Initialisation)
$x_k = \{\}$
$k = 0$
$J_0 = 0$
$S_m = \{f_1, f_2, ..., f_m\}$
$x_k$ is the current feature subset
$m$ is the number of total features
$S$ is the full set of features
while $|x_k| <$ Max feature subset size do
  $f = \text{argmax}_{f \in S_m} \text{eval}(x_k \cup \{f\})$
  $J = \text{eval}(x_k \cup \{f\})$
  if $J > J_k$ then
    $x_{k+1} = x_k \cup \{f\}$
    $S_m = S_m \setminus \{f\}$
    $J_{k+1} = J$
    $k = k + 1$
  else
    break
  end if
end while
Output $x_k$
```

Discussion of the best feature subsets found can be seen in Chapter 7. We note that, of the features explored, the class distributions tend to yield fairly subtle degrees of separation amongst the various types of skin lesion rather than clear cut and easily identifiable clusters within the feature space. The set of features defined and explored in this work is by no means complete or exhaustive and an important direction for future work might look to define further properties which would perhaps yield more representative and distinct features. Even so, the use of depth information does appear to show at least some promise for the purposes of automatic lesion classification.

6. Classification

For the purposes of image classification a standard Bayes classifier is chosen. The Bayes classifier is a classical technique and in this case is a typical representative of supervised machine learning. Here we give a formal definition of the classification technique and provide some discussion towards justifying this choice of classifier.

6.1 Bayes classifier

The technique classifies each previously unseen lesion sample by comparing the relative probabilities of the sample belonging to the different classes in the training set. After finding a suitable feature subset by the method described in Section 5.7.3, we are now able to represent each lesion sample that is to be classified as a feature vector containing values derived from the intensity and range data of the sample. Let this vector be denoted $\mathbf{x} = [f_1, f_2, ..., f_c]^T$. The question we are then interested in for classification becomes:

"given a new sample $\mathbf{x}$, what is $P(class = k|\mathbf{x})$?"
(i.e. *"what is the most likely class for this sample ?"*)

Using Bayes' theorem we are able to evaluate the probability of the sample $\mathbf{x}$ belonging to the class $k \in \{AK, BCC, SCC, ML, SK\}$ *after* we have observed the data in the form of this posterior probability $P(class = k|\mathbf{x})$.

To evaluate this probability, we can use *Bayes rule*

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} \tag{6.1}$$

Which in this case gives

$$P(class = k|\mathbf{x}) = \frac{P(\mathbf{x}|class = k) \cdot P(class = k)}{P(\mathbf{x})} \tag{6.2}$$

6.1.1 Likelihood

Finding the required probabilities involves learning a multivariate Gaussian distribution for each class to find $P(\mathbf{x}|class = k)$ for any datapoint $\mathbf{x}$. The multidimensional Gaussian is given in Equation (6.3) where the D-dimensional vector

μ is the mean and the D × D matrix Σ is the convariance, and $|\Sigma|$ denotes the determinant of Σ.

$$N(\mathbf{x}|\mu, \Sigma) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} \exp\left\{ -\frac{1}{2}(\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu) \right\} \tag{6.3}$$

Finding the best fit for the Gaussian distributions given the training data involves finding the maximum likelihood sample estimates for the two parameters μ and Σ. To determine the parameters from the data, we make the assumption that the data points drawn from this distribution are *independent and identically distributed* (i.i.d.). The maximum likelihood estimates for μ and Σ are then given by

$$\mu = \frac{1}{P} \Sigma_{i=1}^{P} \mathbf{x}^i \tag{6.4}$$

and

$$\Sigma = \frac{1}{P-1} \Sigma_{i=1}^{P} (\mathbf{x}^i - \mu)(\mathbf{x}^i - \mu)^T \tag{6.5}$$

6.1.2 Apriori

Since the denominator on the right hand sides of Equation (6.2) is independent of the class, it remains to find the *apriori* probability $P(class = k)$. The *aprioris* were calculated by simply counting the occurrences of each class in the training data.

The data used in this project is captured during routine diagnosis in the *Department of Dermatology* (cf. Section 3.2). Thus, the dataset may be thought to reflect the *apriori* probabilities of the routine diagnosis in a hospital (note: this is likely not the distribution of lesions observed among the population). We note that using the fraction of observations in the data set for directly estimating the *apriori* probabilities may be subject to over-fitted results for small data sets. Future work might look to develop a full Bayesian treatment for this problem by introducing a prior distribution over the parameter μ, such as a Beta distribution [9]. With a larger data set, it would also be possible to achieve an equal cardinality of each class by selecting an equal number of random samples from each class for the purpose of training the model.

6.1.3 Recognition

Once we have the likelihood and *apriori* probabilities we are able to make use of the Bayes rule relationship (6.6) for classification. We can now obtain the probability we are really interested in, $P(class = k|\mathbf{x})$ and using this can classify each new datapoint $\mathbf{x}$, as the class k that returns the highest posterior probability.

$$\text{posterior} \propto \text{likelihood} \times \text{prior} \tag{6.6}$$

In practice, we take logarithms of the above probabilities as it provides a numerically more stable procedure [5]. Since logarithms are monotonic functions, this will not affect the decision process. Therefore we classify $\mathbf{x}$ as class k where

$$logP(\mathbf{x}|class = k) + logP(class = k) > logP(\mathbf{x}|class = j) + logP(class = j) \ \forall \ j \neq k \tag{6.7}$$

The accuracy of the classifier can then be tested by iterating over each sample in the test data and computing the maximum posterior class using (logged) Bayes rule and comparing the predicted class (largest posterior) with the true class label.

6.1.4 Minimising the expected loss

For this classifier and similar systems used for the analysis of medical images it is noted that the objective may be more complex than simply minimising the number of misclassifications. Considering this skin cancer diagnosis problem in particular, we note that if a patient who does not have a life threatening cancer is incorrectly diagnosed as having such a disease, the consequences may be some patient distress plus the need for further investigations. Conversely, if a patient with a serious, malignant skin cancer is diagnosed as healthy, the result may be premature death due to lack of treatment. Thus the consequences of these two types of mistake can be dramatically different. It would clearly be better to make fewer mistakes of the second kind, even if this was at the expense of making more mistakes of the first kind.

Here we attempt to formalise these issues by introducing a *cost function*, which represents a single, overall measure of loss incurred in taking any of the available decisions or actions [9]. The overall goal then becomes minimising the total cost incurred in taking any of the available classification decisions.

For each new data sample $\mathbf{x}$ of true class k that we assign to class j (where it is possible that $j \neq k$), we incur a level of loss that we denote by L_{kj}, which we define as the k, j element of a *loss matrix*.

By further collaboration with the dermatological expert who provided the original gold standard of clinical diagnoses for the data set, we have defined a set of heuristic cost values for each misdiagnosis based on the following qualitative descriptions.

- A *Extremely severe*: Such a misclasification may result in the death of the patient.
- B *Bad*: Patient death is a possibility, but the problem could be dealt with in a year as easily as now.
- C *Annoying*: The diagnosis slightly overestimates the potential danger of the spot (e.g. a Mole classified as AK or AK classed as BCC).
- D *Extremely annoying*: The true class is likely harmless (i.e. a Mole) but the patient is informed the lesion is something extremely severe.
- E *Indifferent*: The misdiagnosis does not change the outcome or the worry level of the patient.

The proposed heuristic cost values associated with these misclasification descriptions are found in Table 6.1.

Misclasification type	A	B	C	D	E
Proposed cost	10	5	3	5	1

Table 6.1: Heuristic misclasification cost values

From these values we are then able to define a 5 × 5 misclasification loss matrix L tailored[1] to the classes explored in this project as follows:

The decision rule that minimises the expected loss is now the one which assigns each unseen sample $\mathbf{x}$ to the new class j for which the quantity in Equation (6.8) is a minimum. These values are computed easily once we obtain the posterior class probabilities $P(class = k|\mathbf{x})$.

$$\min_{true} \sum_{diag} L_{(true,diag)} P(class = diag|\mathbf{x}) \tag{6.8}$$

We make use of both the standard Bayes classifier decision rule (6.7) and this augmented cost model rule (6.8) for the purpose of skin lesion classification.

[1] Since the classification experiments carried out in this work were performed, an updated loss matrix has been provided with re-evaluated entries for $L_{ML,BCC}$, $L_{ML,SCC}$ and $L_{ML,SK}$

$$\begin{array}{lcc}
 & & \text{Diagnosis} \\
 & & \begin{array}{ccccc} \text{AK} & \text{BCC} & \text{ML} & \text{SCC} & \text{SK} \end{array} \\
\text{True Class} & \begin{array}{l} \text{AK} \\ \text{BCC} \\ \text{ML} \\ \text{SCC} \\ \text{SK} \end{array} & \begin{pmatrix} 0 & C & B & C & B \\ A & 0 & A & E & A \\ C & C & 0 & C & C \\ A & A & A & 0 & A \\ C & D & E & D & 0 \end{pmatrix}
\end{array}$$

Figure 6.1: Loss matrix $L_{(true,diag)}$

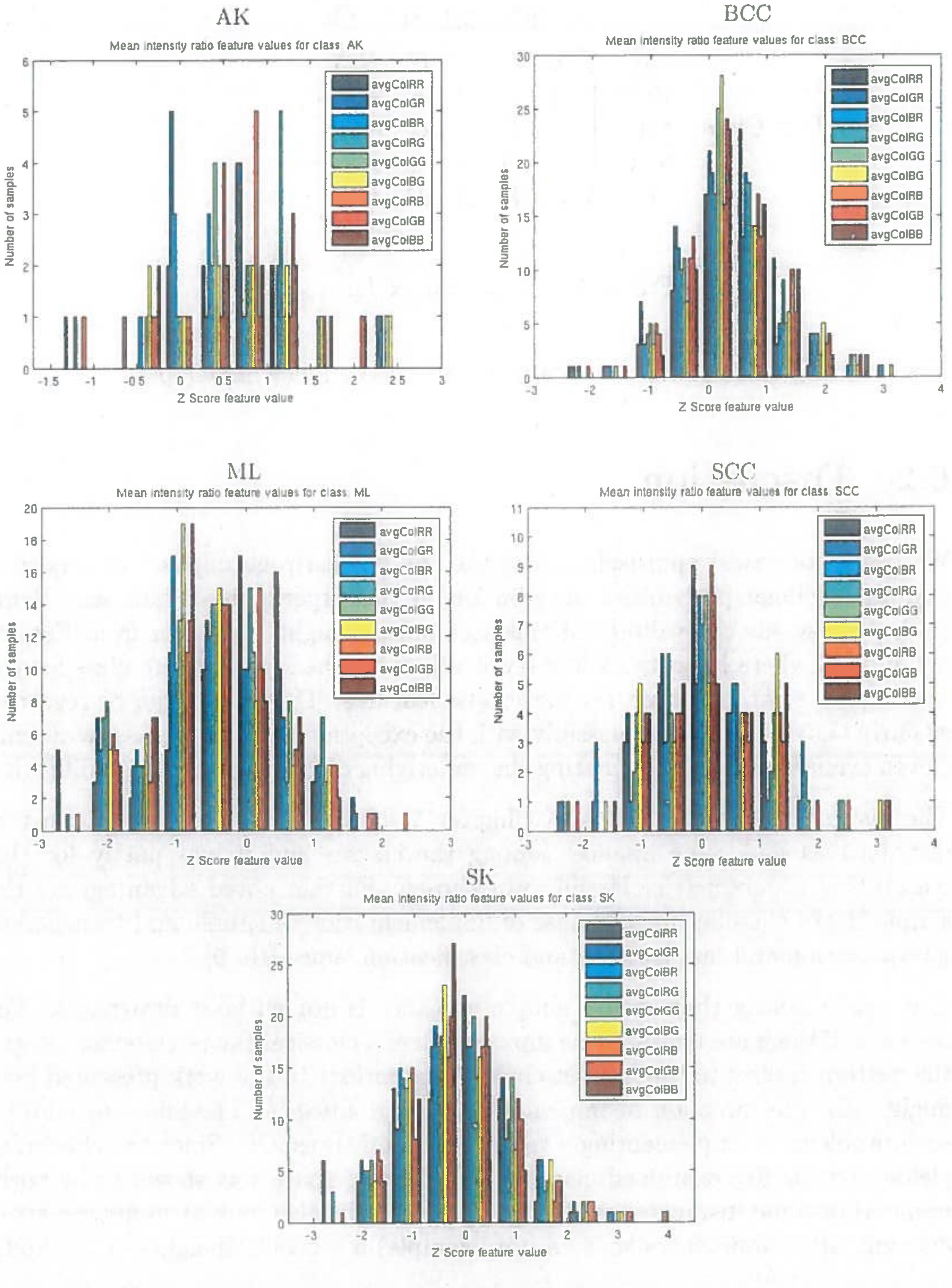
Results using these techniques are discussed in the following chapter.

6.2 Discussion

We use a statistical approach, and make the (typical) assumption of a normal multidimensional probability distribution in the experimental data set. Some evidence towards the validity of this assumption might be drawn from Figures 6.2 and 6.3 where histograms of z-score values can be seen for each class for the *mean spot height* and *mean intensity ratio* features. These plots can be regarded as fairly Gaussian in shape possibly with the exception of the AK class (with only eleven available samples estimating the underlying distribution becomes difficult).

The inspection of the features in Chapter 5 showed a few partially distinctive distributions and some nuances among the classes and it was partly for this reason that a parametric classifier was chosen. Further noted advantages of the simple Bayes classifier include ease of implementation, a simple and transparent interpretation and fast training and classification times [16, 9].

It is worth noting that such a simple classifier is not without drawbacks. For instance, if there are noises in the input set then a classical Bayes classifier assigns the pattern (noise) to the nearest cluster. Extentions to the work presented here might take the direction of implementing more advanced classifiers to address such problems or implementing a rejection cut off threshold. Since the clustering yielded by the five examined classes in the feature space was shown to be fairly minimal in some instances, another direction might also look at nonparametric, discriminative approaches such as (for example) a k-nearest-neighbour method.



AK

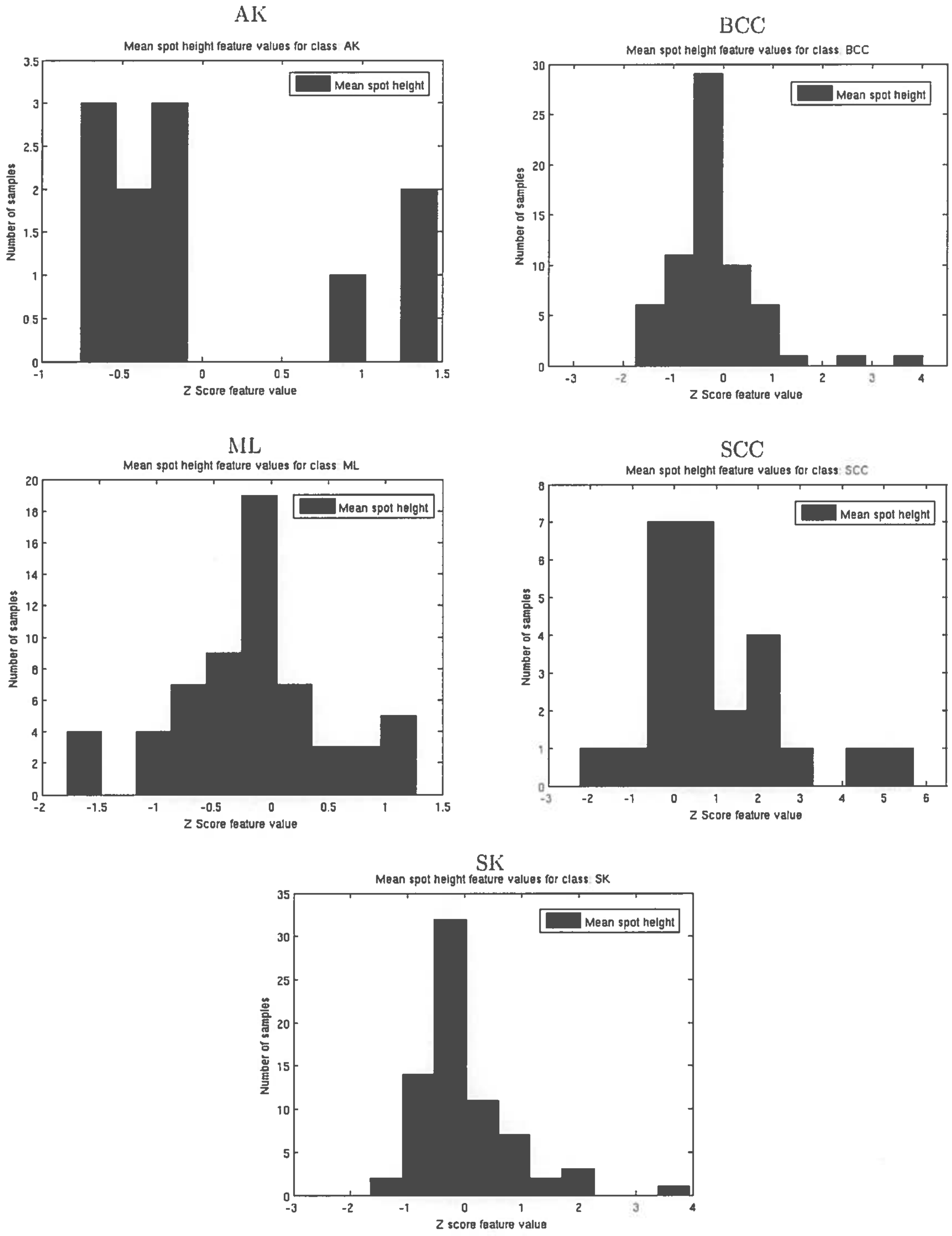
BCC

ML

SCC

SK

Figure 6.2: Class distribution shape for Mean colour intensity ratio feature family



AK

BCC

ML

SCC

SK

Figure 6.3: Class distribution shape for Mean spot height feature

7. Results and evaluation

7.1 Experimental methodology

Obtaining the required evidence for the hypotheses of the project (cf. Section 1.2) involved conducting classification experiments in which the only independent variable was the inclusion or exclusion of depth based features in the system and the dependent observed variable was the resultant classification rates or cost values of the two systems. Multiple classification experiments are used in an effort to average across any random variables outside or beyond our control. These were performed mainly in an attempt to ensure that any difference in accuracy or cost value we are able to measure is due to the presence or absence of depth based features.

7.1.1 Experiment outline

The classification experiments involved splitting the available image samples into disjoint subsets; the *training set* and *testing set* where the training set can be considered the development examples for the systems. Keeping the training and testing sets disjoint ensures that we are measuring how well the systems perform on previously unseen samples. Each experiment involves training two learning systems with identical training sets. Keeping the set constant between systems will help to ensure that the only independent variable in the experiment is the feature set used by the system.

The need to perform multiple classification experiments with a data set of limited size is a problem we note. We make use of cross validation evaluation methods in an effort to 'make the most' of the available data. Due to the fact that the set of lesions deemed usable for experimentation purposes was fairly small, (234 lesion samples over five classes) we make use of this statistical pattern recognition strategy to best exploit the available data for training while still keeping independent data for the performance evaluation. In the following experiments leave-one-out k-fold cross validation was used (i.e. each system is trained on all of the available skin lesions *apart from the one that is to be classified*). This means that 234 separate times, the system is trained on all data except for one point and a prediction is made for that point. This affords us the maximum milage possible from the available data in terms of model training. The main disadvantage experienced with this method was that generating predictions for the entire data set becomes relatively expensive in terms of computation time.

Admittedly the numbers of points in the testing and training sets are ad hoc and additional exploration with a larger data set is likely to be warranted.

Once the systems have been trained, they are then passed the previously unseen testing set images and the classification that the systems make on that sample recorded. Within an experiment, both systems will be tested on the same images. From this record of true and assigned class pairs, we are then able to derive appropriate performance measures such as classification accuracy and minimum cost with which to confirm (or refute) the stated hypotheses. The external gold standard which is required to measure the correctness of the classifications made by the two systems is the hand labelled class of each image which has been provided by a dermatological expert (Professor Jonathan Rees).

7.1.2 Performance metrics

7.1.2.1 Classification accuracy metric

To produce appropriate classification performance measures from the experimental results we compute an *overall classification accuracy* for each system. This is the rate with which it is able to correctly identify *each class* of the provided samples and to take an average of these ratios with respect to the number of classes. This performance measure is defined in Equation (7.1) where N is the number of testing set classes.

$$\text{Overall accuracy} = \frac{1}{\text{N}} \sum_{i=1}^{N} \frac{class_i correct}{class_i total} \tag{7.1}$$

The simpler measure of the ratio of $\frac{\text{total correct}}{\text{total samples}}$ is not used in an effort to avoid the problem of extremely good or bad performance of a highly occurring class weighting the result in a deceptive fashion. For example, as our data set includes a large number of *Seborrheic keratosis* samples, even a naive approach of classifying all test samples as this class would produce a performance better than random class assignment by this simpler metric.

7.1.2.2 Cost function metric

For the second evaluation metric we make use of the loss matrix defined in Section 6.1.4 and define the *cost* of a set of classifications as the sum of the incurred misclasification penalties. In this way a correct classification contributes 0 and the lower overall cost, the better. The cost value for a set of classifications can be easily obtained by multiplying (element-by-element wise) the class predictions represented as a confusion matrix with the loss matrix and summing the resulting entries.

7.2 Experimental results

7.2.1 Feature subset selection

The feature subset selection algorithm described in Section 5.7.3 was used to find the best performing (but not necessarily optimal) feature subsets for both the system constrained to only intensity based features and the system making use of both intensity and depth based data. The algorithm was run for both of the decision rules described in Equations (6.7) and (6.8) in conjunction with the classifier as the criterion function. We note a slight methodological problem here that the criterion function performance is being estimated with the same data set with which classification experiments are performed which may introduce some dependency during feature selection. Ideally, with a larger data set we would leave some of the samples out of the feature selection stage entirely, reserving them completely for classification.

The best performing feature subsets found with respect to the criterion functions are listed in Table 7.1. A numbered feature set index can be found in Appendix. In Figure 7.1 we plot the progressive success rate of the feature selection algorithm using the accuracy criterion function when the selection process has the full range of intensity and depth based features (left) and when it is restricted to intensity features (right). This shows how the rate evolves as more features are selected. We cut off selection after the best ten features are chosen due to the number of available training samples as discussed in Section 5.7. However, it can be noted from the plot that the performance is increasing in a fairly linear fashion and if more data samples were available it might be possible to further increase the accuracy using larger feature subsets (we note however that including too many features runs the risk of overtraining).

Feature set	Constrains on selection	Criterion function	Feature subset
I	Intensity only	Accuracy metric	{10,9,8,3,5,25,4,16,13,7}
II	Intensity and depth	Accuracy metric	{10,9,8,3,**22**,**30**,21,25,15,**26**}
III	Intensity only	Cost metric	{7,6,19,5,9,24,25,13,2,12}
IV	Intensity and depth	Cost metric	{7,6,19,**22**,4,24,8,**1**,**26**,2}

Table 7.1: Selected feature subsets. Numbers corresponding to depth based features are in bold.

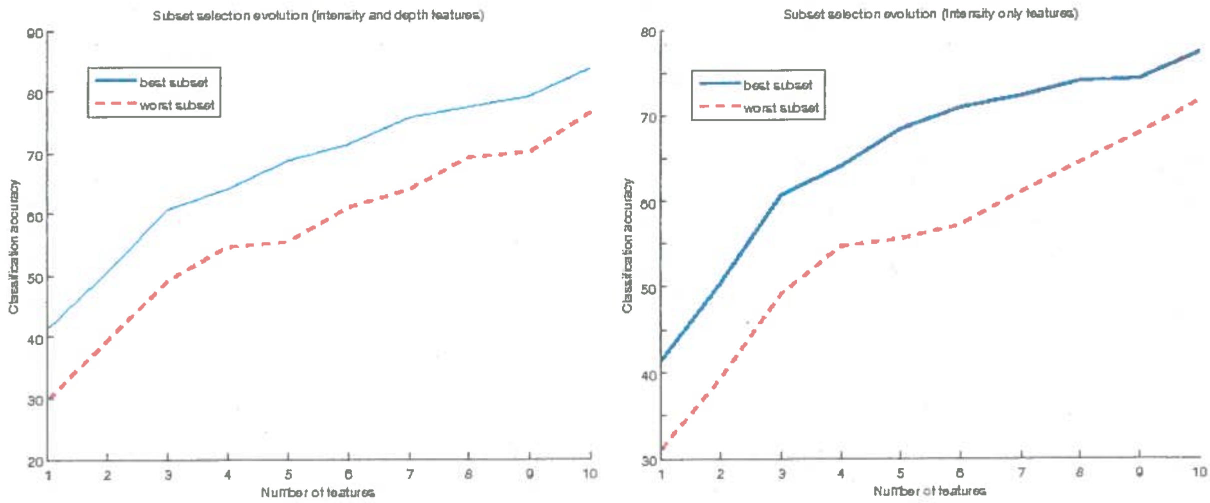


Figure 7.1: Subset evolution comparison

7.2.1.1 Feature selection observations

Contrasting the results of the two selection criterion functions, we can notice from Table 7.1 that the feature sets selected vary widely between criterion, with only features 22 and 26 (peak and pit density in the z-depth channel) common when the algorithm is given the full set of propertries to select from. With only intensity features to select from, it can be seen that five features in total were common between the criterion functions which suggests a little more stability regarding which features are chosen when the available choice is narrower.

Of interest is that in both cases when using the accuracy metric as a criterion function, the *Mean intensity ratio* properties are highly prominent in the selected subset. The intensity only set makes use of seven of the nine properties within this family (10,9,8,3,5,4,7) and the unconstrained set makes use of four (10,9,8,3). The popularity of this family might be partially explained by the fact that this set of features was shown to display a relatively favourable level of class separation in (1D) feature space as discussed in Chapter 5.

Furthermore, with respect to the sets using the accuracy criterion function, we notice that the first four features chosen are the same in both runs, indicating that the initial best four features found using this selection method are intensity based. The features selected were $\frac{\mu_{rS}}{\mu_{bT}}, \frac{\mu_{gS}}{\mu_{bT}}, \frac{\mu_{bS}}{\mu_{bT}}, \frac{\mu_{gS}}{\mu_{rT}}$ from the mean intensity family discussed above (c.f. Section 5.3.1 for detail). After this point, the selected sets fork with the unconstrained set going on to select two depth based features (*peak and pit densities* in the z-depth channel, (**22,26**)) which the constrained run did not have access to. The sets continue to display different features with the unconstrained run picking one further depth based feature (**30**) from the *3D moment invariant*, family.

Similar observations can be made when the cost based criterion is used, where the system able to select from the full feature set selects three properties making use of the depth information channel: *peak and pit density* in the z-depth channel extracted at two filtering levels ($\sigma = 1.0, 2.0$) (features **22,26**) and the *mean spot height* (feature **1**).

It is encouraging that some of the explored features making use of depth information are chosen for the classifier. We concede that the greedy algorithm used here only provides a partial exploration of the feature set space. We highlight that with this algorithm we are not able to delete features from the set once they are added should further additions render them unnecessary or account (to some degree) for the statistical relationship which may exist between the available features[1]. With a more advanced selection algorithm such as Generalised Sequential Forward Selection [6] or those discussed in Section 5.7.3, stronger claims about the usefulness of the depth based features explored might be possible (although compelling proof would likely only come from the exhaustive search of the space).

7.2.2 Classification results

The four feature subsets found in Table 7.1 were used when performing classification experiments on the data set. The results are evaluated by the two performance metrics given in Section 7.1.3. The classification decision rule used in each experiment corresponds to that with which the feature set was chosen (*i.e.* the two models built on a feature set selected by the *cost* based criterion function (sets (III) and (IV)) use the *cost* based decision rule for classification purposes). All classification experiments make use of leave-one-out cross fold validation on the available 234 lesion samples within the data set.

7.2.2.1 Feature sets selected for accuracy

Feature sets (I) and (II) were selected to achieve as good as possible classification accuracy for the five skin lesion categories. Feature set (I) was the best set found from the intensity based features and feature set (II) was the best set found that made use of features from both the intensity and depth information channels. The classification confusion matrices for the results achieved by feature sets (I) and (II) are found in Tables 7.2 and 7.3 respectively.

[1] During the experimentation process it was noted that small alterations to feature definitions often had large knock on effects to the subsets chosen during feature selection. This may suggest the existence of interdependent statistical relationships between the features.

		Diagnosis					
		AK	BCC	ML	SCC	SK	
True	AK	11	0	0	0	0	100%
	BCC	0	55	3	2	5	84.6%
	ML	0	7	49	0	5	80.3%
	SCC	0	9	2	11	3	44%
	SK	0	7	8	1	56	77.7%
Overall accuracy							77.3%
Misclassification cost							306

Table 7.2: Confusion matrix for feature set (I)

		Diagnosis					
		AK	BCC	ML	SCC	SK	
True	AK	11	0	0	0	0	100%
	BCC	0	57	1	4	3	87.6%
	ML	0	4	48	2	7	78.6%
	SCC	0	4	1	19	1	76%
	SK	0	4	8	5	55	76.3%
Overall accuracy							83.7%
Misclassification cost							196

Table 7.3: Confusion matrix for feature set (II)

7.2.2.2 Feature sets selected to minimise cost

Feature sets (III) and (IV) were selected by a criterion to minimise the loss of the function defined in Equation (6.8). This is performed in an attempt to make less misclassifications of a serious nature, possibly at the expense of making more less important mistakes. Feature set (III) was the best available set found for this purpose from the intensity based features alone and feature set (IV) was the best set found that made use of features from both the intensity and depth information channels. The classification confusion matrices for the results achieved by feature sets (III) and (IV) are found in Tables 7.4 and 7.5 respectively.

7.2.2.3 Classification observations

Looking at the classification results of feature set (I) and (II) in Tables 7.2 and 7.3 respectively, we note that the overall achieved accuracy performances, 77.3% and 83.7% can be considered fairly reasonable in both cases. The system which makes use of both depth and intensity based features (set (II)) can be seen to outperform the system using intensity features alone by 6.4% on this particular

		Diagnosis					
		AK	BCC	ML	SCC	SK	
True	AK	11	0	0	0	0	100%
	BCC	0	35	3	25	2	53.8%
	ML	0	3	45	11	2	73.7%
	SCC	0	0	0	24	1	96%
	SK	0	6	16	3	47	65.2%
Overall accuracy							77.7%
Misclassification cost							194

Table 7.4: Confusion matrix for feature set (III)

		Diagnosis					
		AK	BCC	ML	SCC	SK	
True	AK	11	0	0	0	0	100%
	BCC	0	47	3	14	1	72.3%
	ML	0	4	48	4	5	78.6%
	SCC	0	0	1	23	1	92%
	SK	0	6	15	5	46	63.8%
Overall accuracy							81.3%
Misclassification cost							183

Table 7.5: Confusion matrix for feature set (IV)

data set. The question of whether this can be considered significant is addressed in Section 7.2.3.

From the presented results, we can see that using the average class performance as a success metric is favourable to simply building the fraction of the number of all correctly classified lesions over the number of all lesions, since we do not have an equal sample distribution among the classes. If this simpler success estimator was used, it would be possible to fine tune the classification parameters (subsets used) to achieve a better overall performance by improving the accuracy of abundant, less dangerous classes such as ML at the risk of losing performance on the malignant lesion classes.

Taking a closer look at the class specific results in Table 7.2 and 7.3, it can be seen that the two classifiers performed similarly in the classes AK, BCC, ML and SK with both feature sets recognising all AK samples correctly and achieving similar accuracies for the other three classes.

Both systems had some notable trouble recognising the dangerous SCC class, displaying their lowest individual class accuracies. Comparing the two systems with respect to SCC, the system using depth features is able to outperform the one without depth features by a considerable 32% recognition rate. As SCC is a dangerous lesion class, this distinction between the recognition systems is reflected in the higher misclassification penalty (306 vs. 196) awarded to the feature constrained system even though the overall accuracy is similar to that of the system making use of depth features. Typical SCC lesions were noted to have a "crater like" appearance with raised surroundings and a central depression in Chapter 3. The three features utilising depth as a modality in feature set (II) may be at least partially helpful in describing these characteristics and helping to improve the recognition rate for that system.

The most severe problems in these classifications are clearly the cases in which malignant lesions have been categorised as benign spots such as the eight samples in Table 7.2 that really belong to the dangerous BCC class but have been diagnosed as moles and Seborrheic keratosis. In a similar fashion, Table 7.3 exhibits two samples belonging to the malignant class SCC which have been diagnosed as belonging to typically harmless lesion families. An example misclasification can be seen in Figure 7.2 where a true BCC sample was misclassified by the system using feature set (I) as a mole. In this particular example, the system using depth based features was able to correctly identify the BCC class.

The classification systems trained on property sets (III) and (IV) use features selected on the *cost* based criterion and use the *cost* minimising decision rule designed to combat this type of severe misclasification. The classification results for these feature sets are found in Table 7.4 and 7.5.

Comparing the classification results between set (II) and set (IV) (where both sets contain intensity and depth based features and the classifiers differ on the decision rule) it can be noted that there is a sizeable increase in the classification accuracy with which the dangerous SCC class is correctly recognised by the *cost* based approach (from 76% to 92%). This is explained by the cost weighting 'pulling' four SCC samples previously wrongly classed as BCC across the decision boundary. Unfortunately, many true BCC samples that the classifier may not have originally been overly certain about in Table 7.3 have also been pulled across to SCC in Table 7.5, reducing the accuracy here (from 87.6% to 72.3%). This is however deemed a less serious mistake to make as defined by the loss matrix and might therefore be a justifiable compromise for a lesion classifier based in a clinical setting.

In this vein, it is interesting to note that the overall accuracy for the *cost* based approach in Table 7.5 is around 2% lower than the corresponding classifier results in Table 7.3 but, the misclasification cost rating is better (i.e. patient utility may be benefiting more overall). An analogous misleading effect can be seen between the intensity only classifiers in Tables 7.2 and 7.4 where the overall accuracies are very similar but the calculated misclasification *cost* to patients is over one and a half times more when using the original accuracy decision rule!

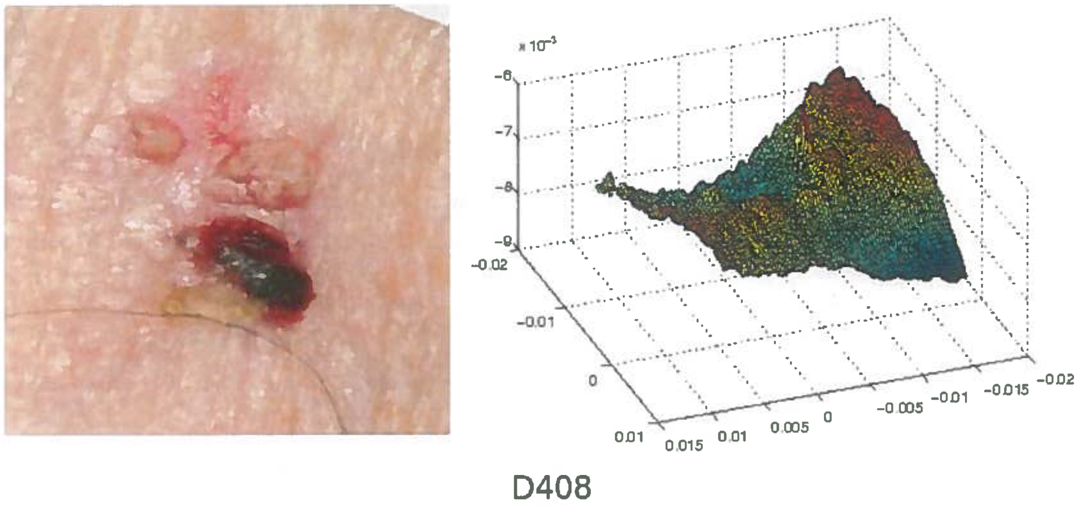


Figure 7.2: Intensity and depth data for sample D408. A BCC lesion misclassified as ML by feature set (I) and correctly classified by set (II)

7.2.3 Statistical analysis

It can be seen in Section 7.2.2 that the classifier making use of depth and intensity based features (set (II)) was able to exhibit a greater overall classification accuracy than that of the classifier using intensity features alone (c.f. Tables 7.2 and 7.3). We now subject this result to basic statistical scrutiny.

We hope to make our classification results amenable to statistical analysis by using common tests from the statistical literature. There are many approximate statistical tests for determining whether one classification system outperforms another on a particular task. One applicable test to verify the statistical significance of our classification results is **McNemar's test** [14]. We now apply this test to the classification results obtained by using feature set (I) (intensity only features) and set (II) (intensity and depth features), where both classifiers made use of the original decision rule found in Equation (6.7).

This test was chosen primarily as it provides a relatively simple yet reportedly accurate measure which can be applied to our dual system classification experiments. Let our base line system be **s1** (making use of only intensity data features) and our system making use of intensity and depth features be **s2**. This will involve keeping track of the following quantities:

# examples misclassified by both **s1** and **s2**	# examples misclassified by **s1** but not by **s2**
# examples misclassified by **s2** but not by **s1**	# examples misclassified by neither **s1**,**s2**

We now abbreviate these quantities with the following identity:

n_{00}	n_{01}
n_{10}	n_{11}

The sum of these four quantities is the total number of examples in the test set. We can construct a null hypothesis that **the two systems have the same error rate**. McNemar's test is based on a χ^2 test and essentially computes a goodness of fit that compares the distribution of counts expected under the null hypothesis to the observed counts. The expected counts under the null hypothesis are:

n_{00}	$(n_{01} + n_{10})/2$
$(n_{01} + n_{10})/2$	n_{11}

The test then specifies the following statistic (Equation (7.2)) which is distributed (approximately) as χ^2 with 1 degree of freedom.

$$\frac{(|n_{01} - n_{10}| - 1)^2}{n_{01} + n_{10}} \tag{7.2}$$

If the null hypothesis is correct at the 95% confidence level, then the probability that this quantity is greater than $\chi^2_{1,0.95} = 3.841459$ is less than 0.05. So we may reject the null hypothesis in favour of the hypothesis that the two systems have different performance when trained on this particular training set.

7.2.3.1 Statistical results

The classifier results being tested (Tables 7.2 and 7.3) yielded the following observed counts for McNemar's test.

25	19
27	163

We can therefore compute the statistic value for the classifications made by the two recognition systems as:

$$\frac{(|19 - 27| - 1)^2}{19 + 27} = \frac{49}{46} \approx 1.065 \tag{7.3}$$

We concede that this χ^2 value is not above the required level to be significant $p < 0.05$. We cannot reject the null hypothesis at this significance level. As $\chi^2_{1,0.698} = 1.065$ we are however able to reject the null hypothesis in favour of the hypothesis that the two classifiers have statistically different performances when trained and tested on this data set at the lower significance level of $p < 0.302$ (i.e. there is less than 30% chance that this result arose by statistical variation).

In this respect we have provided some evidence to support our initial hypothesis that a classification system using a combination of intensity and depth based image features is more successful at the task of classifying skin lesions than a system which uses intensity based features alone.

8. Conclusion

8.1 Discussion and summary

This work investigated the benefits of using surface shape and depth information for automated skin cancer classification. We used a novel application of stereo photogrammetry techniques to obtain surface shape information from a collection of skin lesion samples.

The capture equipment allows for intensity and depth information to be acquired in a fast, noninvasive and accurate manner which yields a favourable technique for the specialised task of recording skin lesion information. The particular lesion data provided for this work originated from an independent clinical environment which proved to be a challenging source, exhibiting variations in the size of lesion classes, quality of the data captured and distributions of data properties amongst samples.

These characteristics of the data highlighted the need for accurate initial capture and further processing of the information channels to enhance features. In particular, the depth data channel was fitted to a surface to best preserve the presence of interesting local texture, with some success. Sometimes this technique was observed to fail, with global surface structure such as body parts or hair follicles obscuring the area of interest. Depth information was successfully used at this processing stage as input to visualisation techniques which were able to provide useful information about the quality of the available data, highlighting a use outwith classification.

The features used in the integrated system were designed to some extent, to reflect the parameters used in medical diagnosis. The distributions of the data samples in the examined feature spaces tended to exhibit nuances and subtle degrees of separation rather than distinct clusters. This trend was seen to be fairly universal across intensity and depth information channels.

By application of feature selection strategies the number of features used for classification was reduced from 30 to 10 parameters. It was noted that the classification success of the system increases in a fairly linear fashion as more features are added to the selected subset. This trend suggested that if more data points were available, further increases in classification performance might be possible by making use of larger feature subsets.

A simple statistical approach to classification was implemented. The overall accuracy results of 77.3% and 83.7% for systems using intensity only and both depth and intensity based feature sets depict a reasonable performance for an automated system on a challenging data set. It was also shown that changing the question being asked by the classifier decision rule could notably reduce the number of dangerous misclassifications being made by the system, without sacrificing the overall classification accuracy.

By subjecting classification results to statistical testing, we were able to produce evidence at a reduced confidence level in support of our original hypothesis that the use of range data is beneficial to automated skin lesion diagnosis. In summary, depth information shows potential as a useful modality for skin lesion classification especially if we can guarantee accurate capture conditions.

8.2 Further work

This work has only looked at a small set of data points and feature types. Further exploratory investigation is warrented in a number of directions. In order to achieve better classification results, new parameters should be derived and included in the feature set. Extracting richer and more advanced properties that are able to exploit the available depth information might yield a good starting point for future work. New features might take the form of more complex treatments of local roughness, curvature in the surface or global structure measures. One interesting avenue to follow that might yield good results would be the *combination* of 2D and 3D based features, which was not explored here.

As noted previously, more powerful feature subset selection techniques might be applied to the problem of feature selection in this domain to combat the suspected problem of feature interdependency when selecting subsets. With respect to exhaustively searching the space, one route we were not able to fully explore in this project was the possibility of applying distributed computing power to the problem. The need to reduce the dimensionality of the feature space might also be attacked from a different angle, such as Principal Component Analysis.

More advanced and exotic classification algorithms could certainly be applied to this problem but as noted in the previous work survey, diversity in this respect tends to make comparing related research difficult. Of more use would be a coordinated effort to collect and distribute a shared set of high quality depth and intensity image data upon which, standard classification techniques might be agreed.

As the calibre of camera hardware improves and capture techniques mature, higher quality depth based data is likely to become available. Combining richer data with some of the proposed recommendations may have the potential to greatly improve the ability to automatically recognise skin lesion structure and form and in turn, help to fully automate the diagnosis process.

9. Appendix

Feature Index

1. Mean spot height
2. $\frac{\mu_{rS}}{\mu_{rT}}$
3. $\frac{\mu_{gS}}{\mu_{rT}}$
4. $\frac{\mu_{bS}}{\mu_{rT}}$
5. $\frac{\mu_{rS}}{\mu_{gT}}$
6. $\frac{\mu_{gS}}{\mu_{gT}}$
7. $\frac{\mu_{bS}}{\mu_{gT}}$
8. $\frac{\mu_{rS}}{\mu_{bT}}$
9. $\frac{\mu_{gS}}{\mu_{bT}}$
10. $\frac{\mu_{bS}}{\mu_{bT}}$
11. $\frac{\sigma_S^r}{\sigma_T^r}$
12. $\frac{\sigma_S^g}{\sigma_T^g}$
13. $\frac{\sigma_S^b}{\sigma_T^b}$
14. $\frac{\sigma_S^z}{\sigma_T^z}$
15. $Peakandpitdensity_r(\sigma = 0.5)$
16. $Peakandpitdensity_g(\sigma = 0.5)$
17. $Peakandpitdensity_b(\sigma = 0.5)$
18. $Peakandpitdensity_z(\sigma = 0.5)$
19. $Peakandpitdensity_r(\sigma = 1.0)$
20. $Peakandpitdensity_g(\sigma = 1.0)$
21. $Peakandpitdensity_b(\sigma = 1.0)$
22. $Peakandpitdensity_z(\sigma = 1.0)$
23. $Peakandpitdensity_r(\sigma = 2.0)$

24. $Peakandpitdensity_g(\sigma = 2.0)$
25. $Peakandpitdensity_b(\sigma = 2.0)$
26. $Peakandpitdensity_z(\sigma = 2.0)$
27. Δ spot diameter[†]
28. J_1
29. J_2
30. J_3

[†] An experimental feature which was not used in the evaluation experiments.

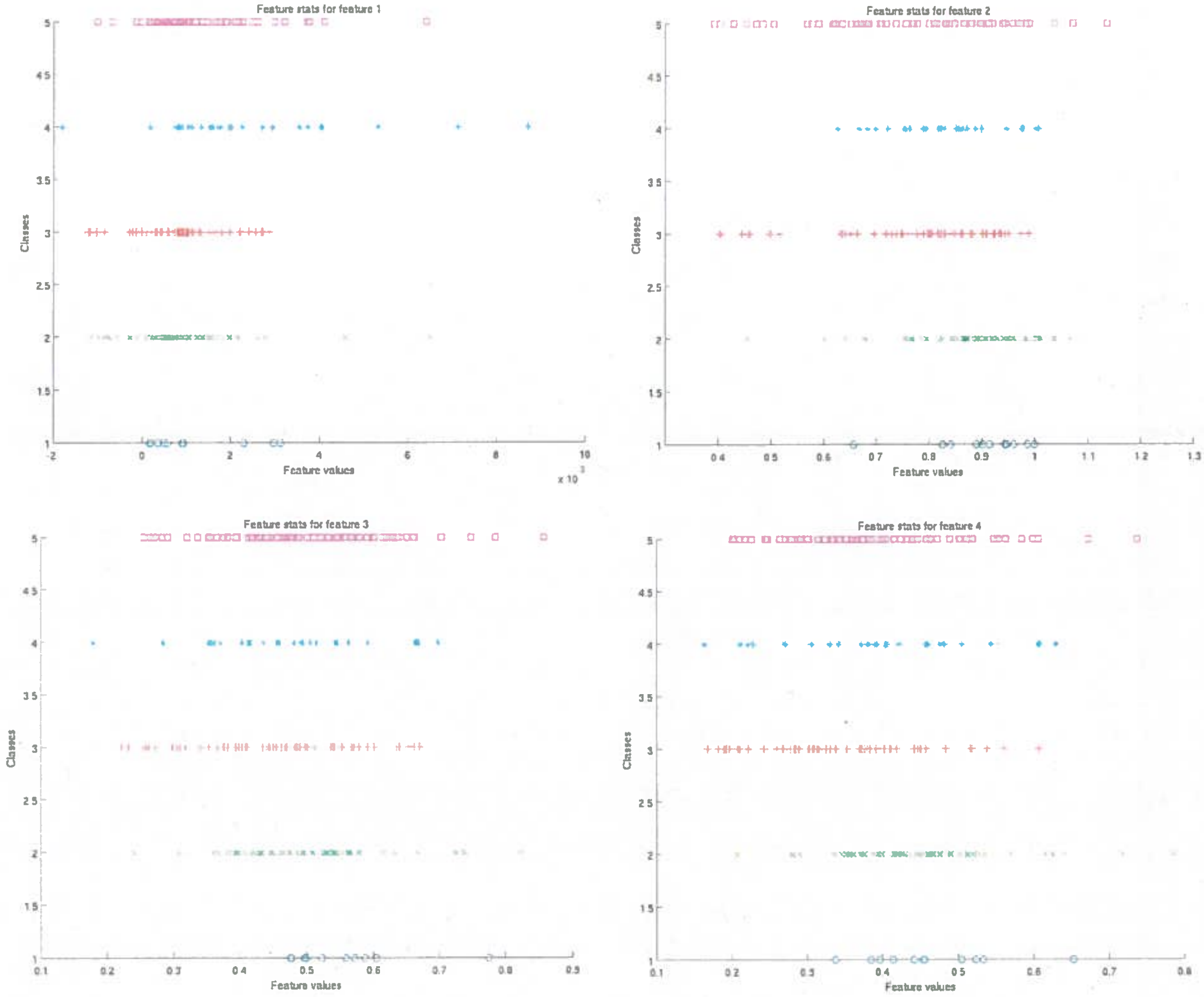


Figure 9.1: Class distributions for features 1-4

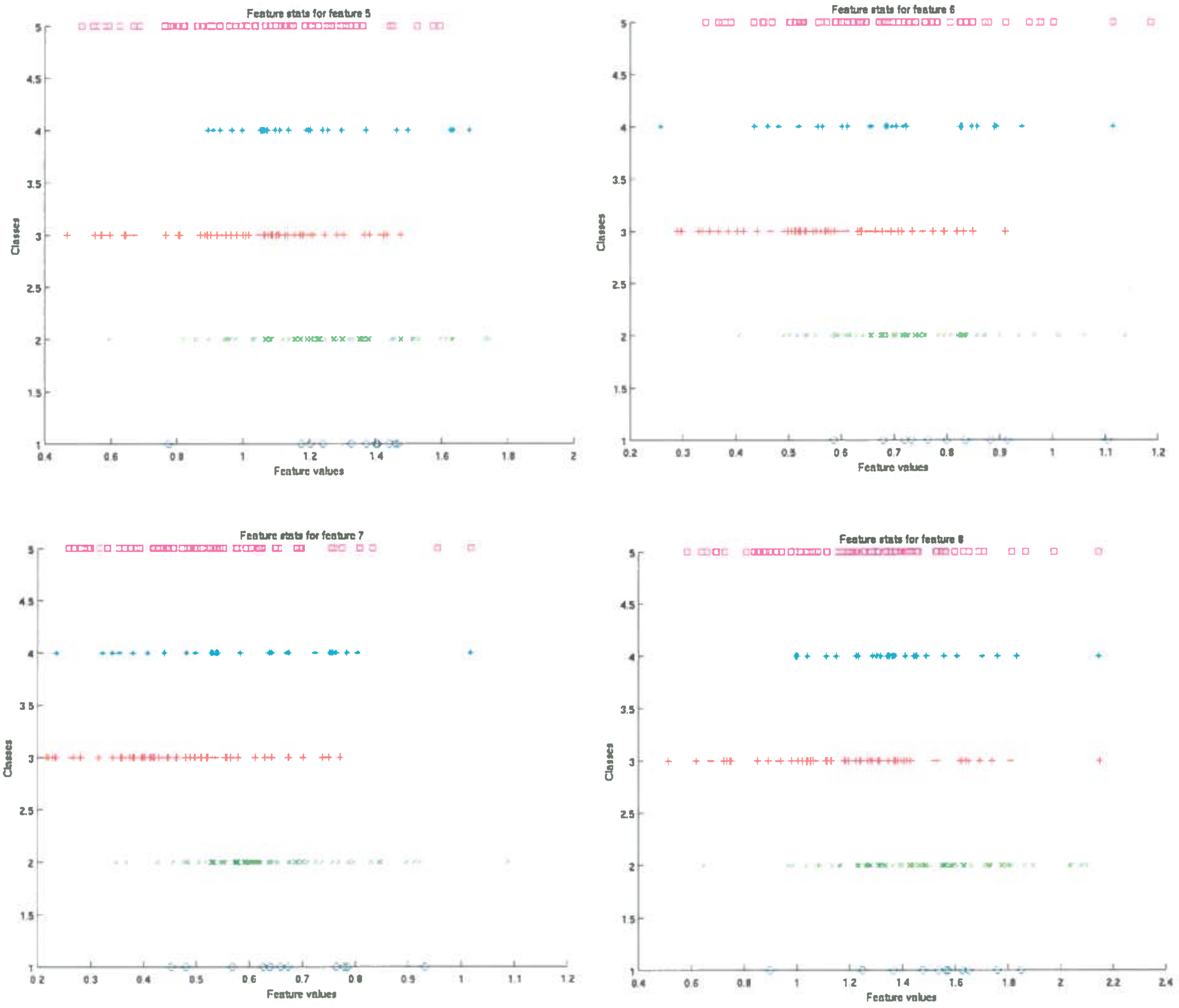


Figure 9.2: Class distributions for features 5-8

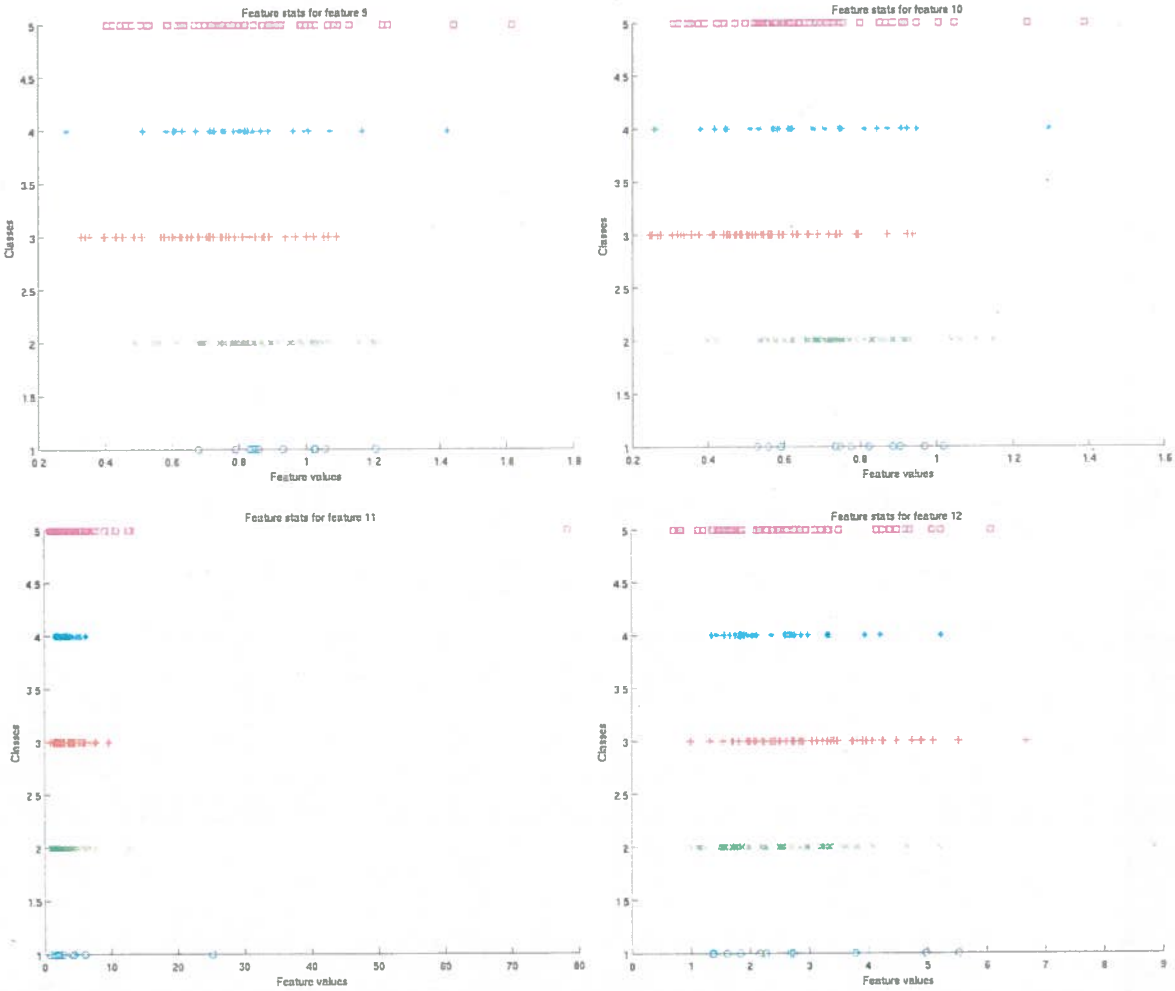


Figure 9.3: Class distributions for features 9-12

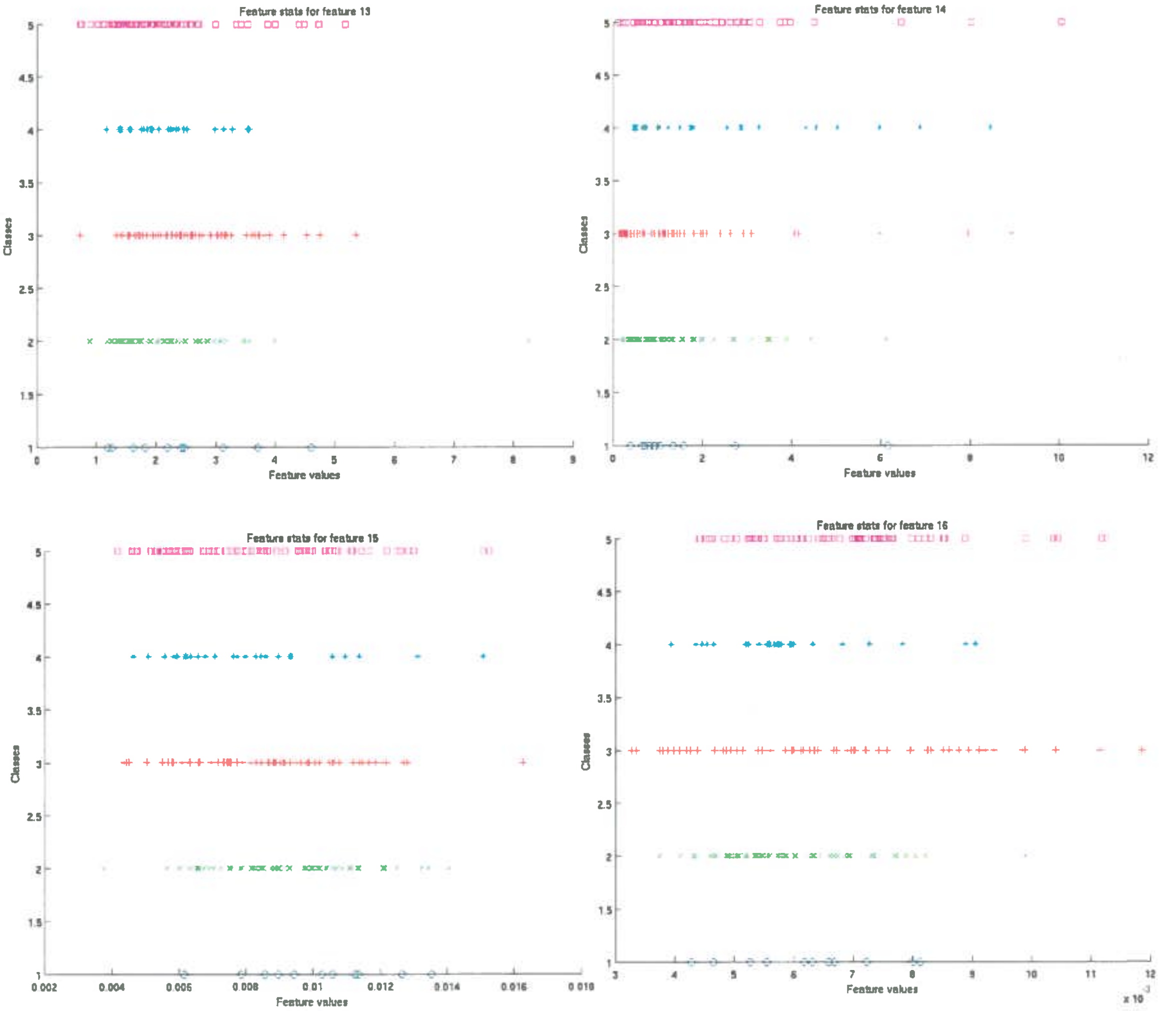


Figure 9.4: Class distributions for features 13-16

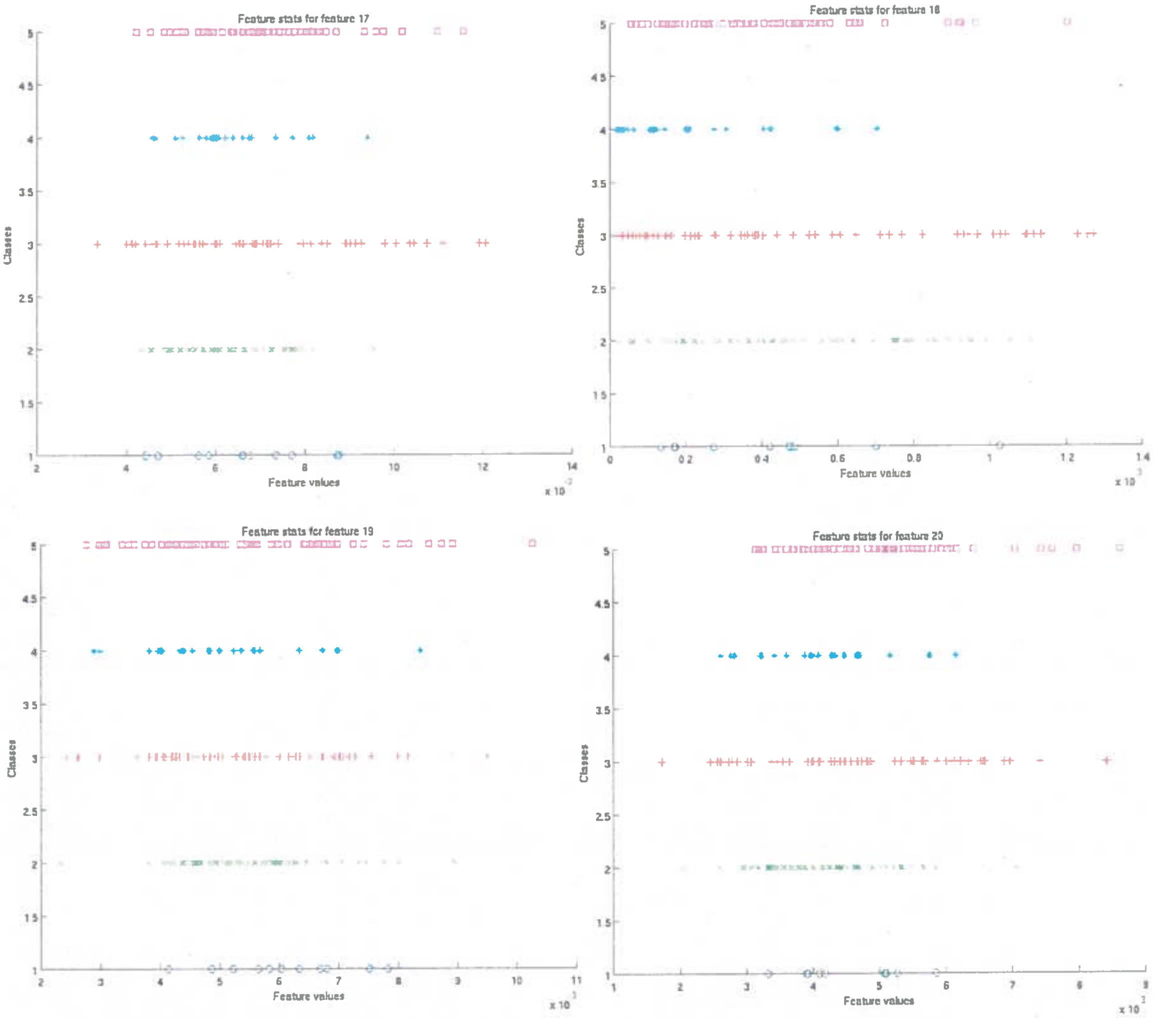


Figure 9.5: Class distributions for features 17-20

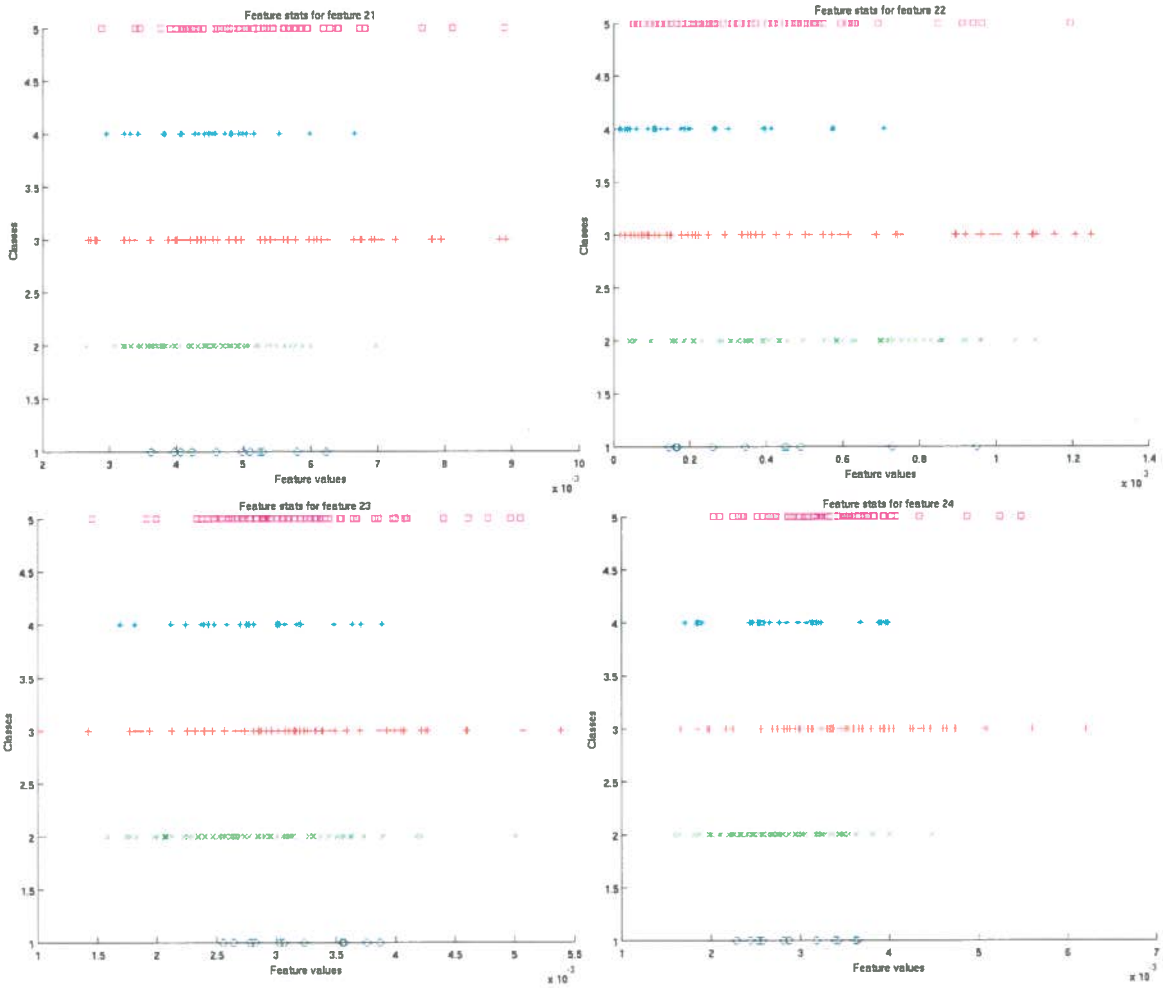


Figure 9.6: Class distributions for features 21-24

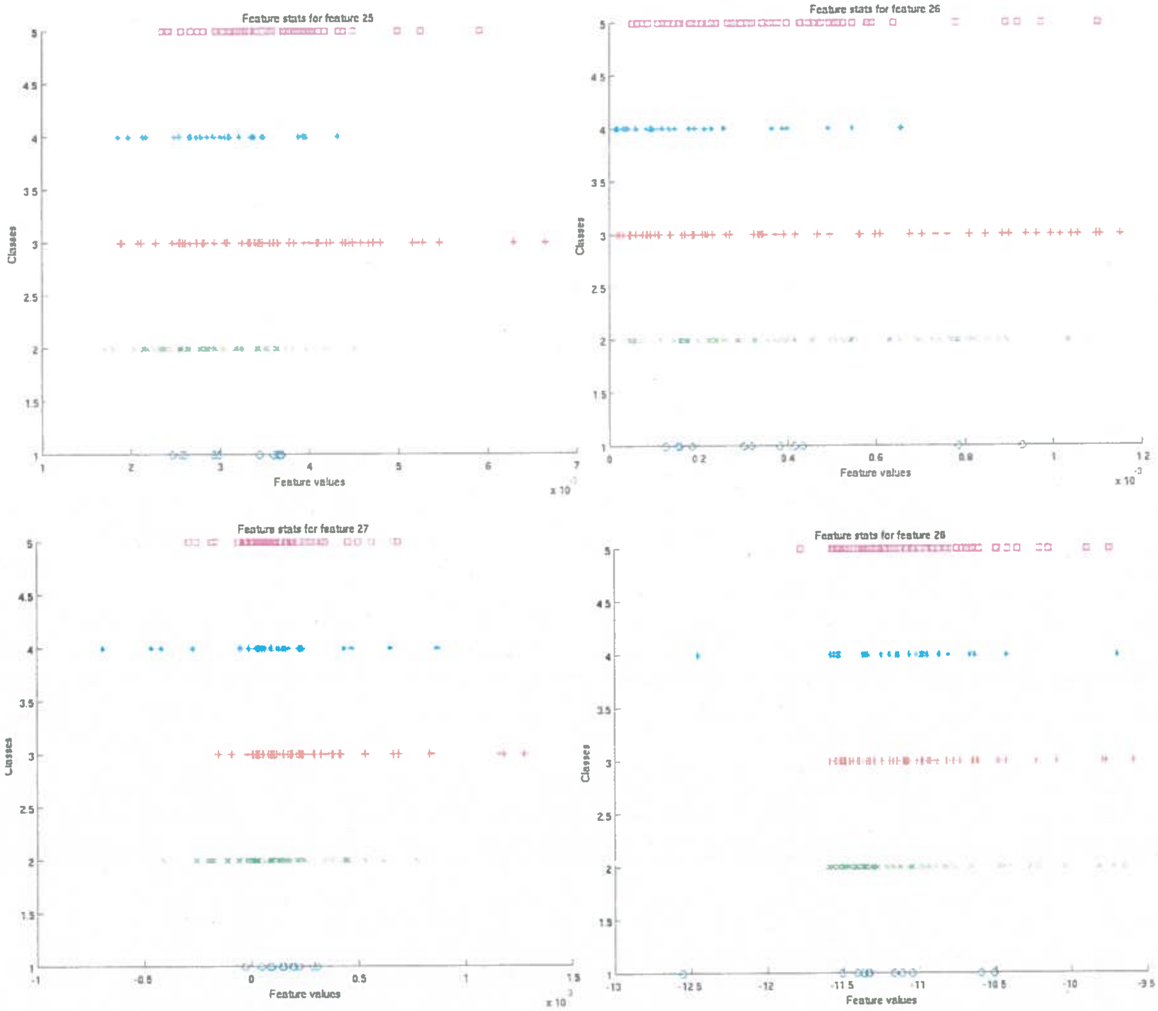


Figure 9.7: Class distributions for features 25-28

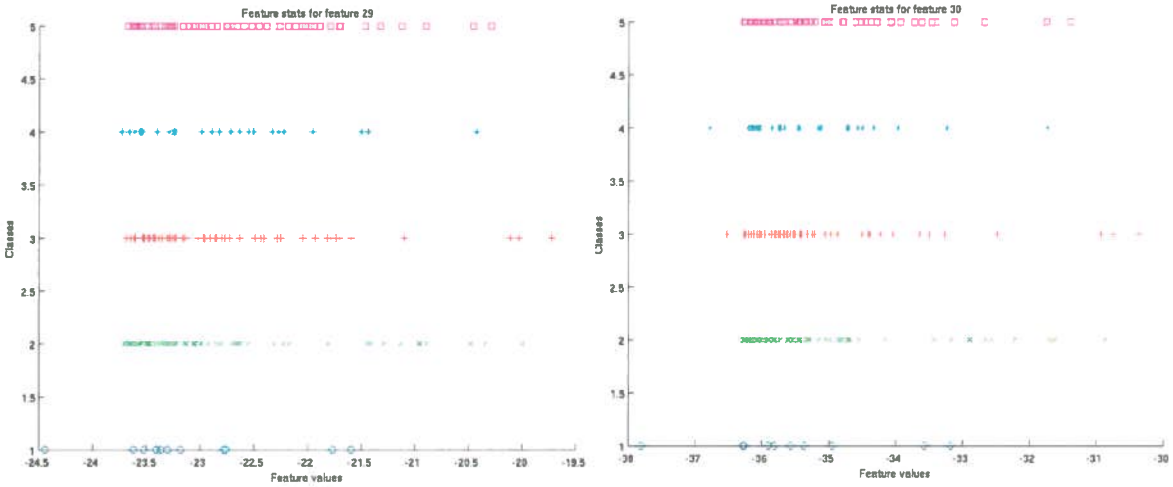


Figure 9.8: Class distributions for features 29-30

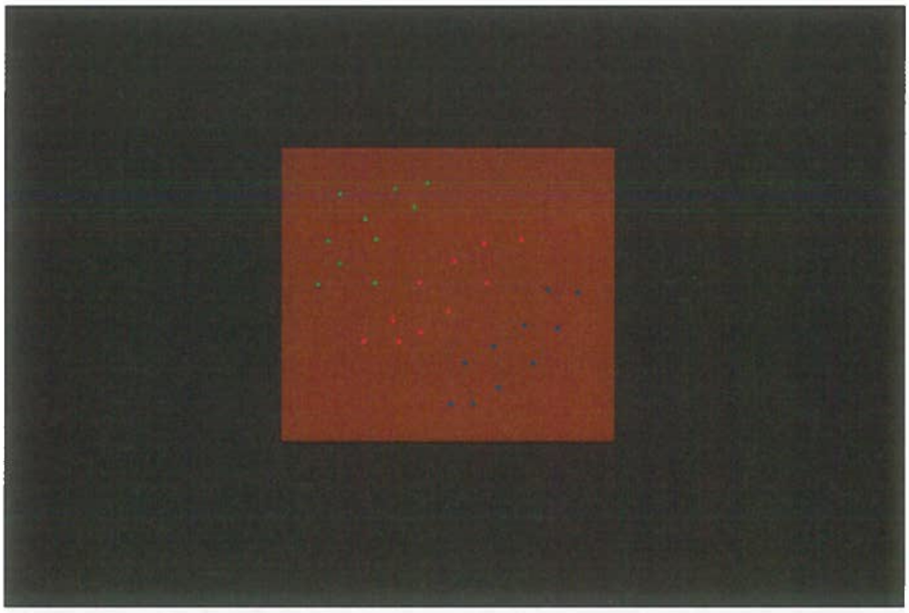


Figure 9.9: Synthetic intensity data for peaks and pits verification

Bibliography

[1] http://www.di3d.com/page_technology.html Accessed: 20/02/08.

[2] http://www.dcs.gla.ac.uk/publications/PAPERS/7029/Scan2003_xju2.jps.PDF Accessed: 4/12/07.

[3] www.skincancer.org Accessed 22/01/08.

[4] R. P. W. Duin A. Jain and J. Mao. Statistical pattern recognition: a review. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22:4–37, 2000.

[5] David Barber Amos Storkey. Learning from Data, Density Estimation: Gaussians, Accessed: 11/02/08. http://www.inf.ed.ac.uk/teaching/courses/lfd/lectures/lfd_2005_density.pdf.

[6] Alex M. Andrew. Statistical Pattern Recognition, London (Cambridge University Press, New York, for USA). *Robotica*, 18(2):219–223, 2000.

[7] Dhawan AP. An expert system for the early detection of melanoma using knowledge-based image analysis. *Anal Quant Cytol Histol*, 10:405–416, 1988.

[8] Breen E et al. Bischof L, Talbot H. An automated melanoma diagnosis system. Technical report, University of Ballarat / University of Technology, Sydney. Presented at the Research Workshop on Automated Medical Image Analysis. 1998.

[9] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, August 2006.

[10] Marco Callieri, Paolo Cignoni, Paolo Pingi, Roberto Scopigno, M. Coluccia, G. Gaggio, and Maria Novella Romanelli. Derma: Monitoring the evolution of skin lesions with a 3D system. In *VMV*, pages 167–174, 2003.

[11] Bufalino R et al Cascinelli N, Ferrario M. Results obtained by using a computerized image analysis system designed as an aid to diagnosis of cutaneous melanoma. *Melanoma Res*, 2:163–170, 1992.

[12] M. Dash and H. Liu. Feature selection for classification, 1997.

[13] Kittler J Devijver PA. *Pattern Recognition: A Statistical Approach*. Prentice-Hall, 1982.

[14] Thomas G. Dietterich. Approximate statistical test for comparing supervised classification learning algorithms. *Neural Computation*, 10(7):1895–1923, 1998.

[15] B. Fang, W. Hsu, and M. Lee. Tumor cell identification using feature rules, 2002.

[16] Bob Fisher. Introduction to Vision and Robotics: Lecture note, Accessed: 23/12/07. http://www.inf.ed.ac.uk/teaching/courses/ivr/lectures/ivr6hand.pdf.

[17] R H Barbour G R Day. Automated melanoma diagnosis: where are we at? *Skin Research and Technology*, 6:1–5, 2000.

[18] Pfitzner J O'Rourke M Knight N. Green A, Martin N. Computer image analysis in the diagnosis of melanoma. *J Am Acad Dermatol*, 31:958–964, 1994.

[19] R Rohrer E Wilding M Binder H Kittler H Ganster, A Pinz. Automated melanoma recognition. *IEEE Transactions on Medical Imaging*, 20:234–239, 2001.

[20] Morris Smith J. Hall P, Claridge E. Computer screening for early detection of melanoma - is there a future? *Br J Dermatol*, 132:325–338, 1995.

[21] M. K. Hu. Visual pattern recognition by moment invariants. *IRE Trans. Info. Theory*, IT-8:179–187, 1962.

[22] Kohavi R. John, G.H. and K. Pfleger. Irrelevant features and subset selection problem. *Proceedings of the Eleventh International Conference on Machine Learning*, pages 121–129, 1994.

[23] K. Kira and L.A. Rendell. The feature selection problem: Traditional methods and a new algorithm. *Proceedings of Ninth National Conference on Artificial Intelligence*, pages 129–134, 1992.

[24] D. Koller and M. Sahamni. Toward optimal feature selection. *In Proceedings of International Conference on Machine Learning*, 1996.

[25] Martin Kreutz, Maik Anschutz, Stefan Gehlen, Thorsten Grunendick, and Klaus Hoffmann. Automated diagnosis of skin cancer: Using digital image processing and mixture-of-experts. In *Bildverarbeitung fur die Medizin*, pages 357–361, 2001.

[26] Tim Lukins. Micro-scale stereo photogrammetry of skin lesions for depth and colour classification, 2005.

[27] Randy H. Moss Murali Anantha and William V. Stoecker. Detection of pigment network in dermatoscopy images using texture analysis. *Computerized Medical Imaging and GraphicsVolume*, 5(28):225–234, 2004.

[28] P.M. Narendra and K Fukunaga. A branch and bound algorithm for feature selection. *IEEE Transactions on Computers*, 26:917–922, 1977.

[29] J. Novovicov P. Pudil and J. Kittler. Floating search methods in feature selection. *Pattern Recognigion Letters*, 15:1119–1125, 1994.

[30] D. S. Rigel R. J. Friedman and A. W. Kopf. Early detection of malignant melanoma: the role of physician examination and self-examination of the skin. pages 130–151, 1995.

[31] Hriday Ravindranath. Skin Spot Classification using 3D data. *MSc Thesis, School of Informatics, University of Edinburgh*, 2006.

[32] John C. Russ. *The image processing handbook (3rd ed.)*. CRC Press, Inc., Boca Raton, FL, USA, 1999.

[33] F. A. Sadjadi and E. L Hall. Three-dimensional moment invariants. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2(2):127–136, 1980.

[34] Pepe P. Seidenari S, Pellacani G. Digital videomicroscopy improves diagnostic accuracy for melanoma. *J Am Acad Dermatol*, 39:175–1181, 1998.

[35] S. Sigurdsson, P. A. Philipsen, L. K. Hansen, J. Larsen, M. Gniadecka, and H. C. Wulf. Detection of skin cancer by classification of raman spectra. *IEEE Transactions on Biomedical Engineering*, 51(10):1784 – 1793, 2004.

[36] Moss RH. Stoecker WV. Editorial: Digital imaging in dermatology. *Comput Med Imag Graphics*, 16:145–150, 1992.

[37] Jonathan Rees Colin Urquhart Tim Lukins, Bob Fisher. Micro scale photogrammetry of skin lesions: Poster demonstration, 2005.