Classification of Microcalcifications in Digitised Mammograms

Dani Kramer

A dissertation submitted to the Faculty of Engineering, University of the Witwatersrand, Johannesburg, in fulfilment of the requirements for the degree of Master of Science in Engineering.

Johannesburg, April 1999

Declaration

I declare that this dissertation is my own, unaided work, except where otherwise acknowledged. It is being submitted for the degree of Master of Science in Engineering in the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination in any other university.

Signed this 4th day of April 1999

Dani Kramer.

Abstract

In this investigation a number of image texture analysis techniques for the classification of microcalcifications in digitised mammograms are presented. Microcalcifications are often an early indication of breast cancer, and computer-aided diagnostic techniques are capable of improving diagnostic accuracy. Three categories of image texture features are extracted from regions of interest surrounding clusters of microcalcifications. These comprise a set of statistical texture features based on the co-occurrence matrix, a set of wavelet-based texture signatures and a proposed third set of texture features. This set, referred to as multiscale statistical texture features, is based on a combination of the other two approaches to texture analysis. The multiscale statistical texture features outperform the other types of texture features in tests using two separate datasets and a k-nn classifier for classification. Improved classification accuracy is also achieved using an artificial neural network for classification.

Acknowledgements

I am grateful to my supervisor, Dr Farzin Aghdasi, my colleagues Richard Pfisterer and Aydin Alaylioglu, Jessica Heynis and my family, for their assistance and support during the preparation of this dissertation.

I also wish to acknowledge the sources of the mammographic databases used for this study. These were provided by courtesy of the National Expert and Training Centre for Breast Cancer Screening and the Department of Radiology at the University of Nijmegen, the Netherlands, and the Lawrence Livermore National Laboratories (LLNL) and The University of California at San Francisco (UCSF) Radiology Department, e-mail: mammo-db-help@llnl.gov.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Problem Description

Breast cancer is one of the leading causes of cancer-related death amongst middle-aged and older women. Unfortunately, to date, primary prevention of the disease is not possible as its cause is not yet fully understood. Treatment, however is very successful if the cancer is detected in the early stages of development. The only promising method of treatment therefore is to detect the disease at the earliest possible stage. As a result, the most effective approach of combating breast cancer is through population-wide, interval screening programs of asymptomatic women [1].

X-ray mammography is currently the best method for the early detection of breast cancer. Mammograms enable radiologists to visualise nonpalpable tumours and microcalcifications, both early indications of breast cancer. Microcalcifications are minute deposits of calcium that often indicate an active secretory process of tumour cells. However, although mammography is currently the best method for the early detection of breast cancer, 10%–30% of women who have breast cancer and undergo mammography have negative mammograms [2]. In addition, in approximately two-thirds of these false-negative mammograms, the cancer is evident in retrospect [3]. It has also been shown that double reading of mammographic films, by two different radiologists, increases the detection sensitivity by between 5% and 15% [4]. Therefore in order to increase the sensitivity of mammographic screening programs, many computer-aided diagnostic techniques have been developed as a tool to aid radiologists in the detection process.

In this application, computer-aided diagnosis (CAD) may be defined as the diagnosis made by a radiologist who takes into consideration the results of a computerised analysis of radiographic images, and uses this analysis as a "second opinion" for the detection of lesions and during the diagnostic process. The final diagnosis is made by the radiologist. Since mammography is routinely implemented in wide-scale screening programs where radiologists do not detect all cancers that are visible in retrospect, and due to the effectiveness of double-reading, it is expected that the efficiency and effectiveness of screening for breast cancer can be increased using CAD techniques [3].

Once a suspicious abnormality has been detected by a radiologist, he or she must visually extract various radiographic characteristics, and use these characteristics to distinguish between benign and malignant abnormal lesions in order to decide what course of action should be taken, i.e. should the patient return to the interval screening program, return for a follow up examination or return for a surgical biopsy. However, although general rules exist for the differentiation between benign and malignant breast lesions, most suspicious lesions are referred for surgical biopsy examination. Of these only between 15% and 30% are actually cancerous [5]. Thus, it is also hoped that CAD techniques may be successfully applied to the task of differentiating between benign and malignant abnormal lesions.

The purpose of this work is to investigate and develop techniques suitable for distinguishing between benign and malignant microcalcifications in digitised mammograms. The possibility of developing a non-invasive method of predicting the outcome of a biopsy would represent an aid to radiologists in their classification of abnormal lesions, and therefore reduce the high percentage of false biopsy examinations. Improvement in the classification accuracy would decrease patient trauma, scarring, and the implicit surgical complications caused by unnecessary surgical biopsies. Furthermore, it would reduce the high cost associated with the large volume of unnecessary biopsies performed in wide-scale mammographic screening programs.

In this analysis an attempt is made to characterise microcalcifications as benign or malignant based on information available in the radiographic view, thus the classification system utilises digitised mammograms for the analysis.

In a clinical environment, a computerised classification system would be used as a "second opinion", for use in wide-scale screening programs. The system is an especially useful tool for the inexperienced mammographer or general radiologist.

This work focuses on the development of a system to classify microcalcifications as benign or malignant and is therefore dependant on either a manual or an automatic microcalcification detection scheme.

1.2 Proposed Solution

Many classification systems have been developed that are capable of differentiating between benign and malignant microcalcifications based on information contained in the radiographic view. These approaches will be discussed in detail in Chapter 2. Most of these techniques first segment the digitised grey-level images into binary mask images containing the extracted microcalcifications. Features are then extracted from these images. For example, the shape, size, orientation and number of microcalcifications are used as discriminatory features for classification. The methods thus rely on the perfect segmentation of the microcalcifications. This is often unreliable, especially in difficult to diagnose cases. In other systems, the manual intervention by a radiologist is required in order to extract distinguishing features directly from the mammogram, which are then used as a basis for classification.

In this work a texture-based approach for the classification of microcalcifications is developed. The conjecture is that there exists some form of textural information in the areas surrounding the clusters of microcalcifications, and that this information may be utilised as a basis for the analysis of the microcalcifications.

The microcalcifications themselves are merely an indication of possible tumour activity, and texture analysis techniques which attempt to characterise the underlying tissue surrounding the area containing the microcalcification, may therefore prove useful in the diagnostic process.

The classification system relies on a set of texture features computed directly from the regions of interest (ROI) containing the microcalcifications. This set of texture features are then used to discriminate between benign and malignant microcalcification areas. The feature set is extracted directly from the ROI containing the microcalcification. There is therefore no need for precise segmentation algorithms. The classification system only requires the coordinates of the suspicious lesion as an input. It should also be noted that in the diagnosis of a cluster of microcalcifications, the radiologist examines the physical arrangement of the microcalcifications. This system thus provides additional information as it analyses the texture surrounding the microcalcifications, not the microcalcifications themselves.

This study compares the suitability of a number of existing approaches for the analysis of image texture to the problem of the classification of the texture surrounding microcalcifications in digitised mammograms. The inclusion of a more generalised approach to image texture analysis is also proposed. The first group of texture features used are based on the texture analysis techniques proposed by Haralick [6], which constitute a statistical analysis of the grey-levels in an image. The second group of texture features are wavelet-based multiscale texture features, similar to those used by Dhawan [7] and Kocur [8]. The proposed third group of texture features are based on a combination of the above texture features, i.e. the statistical and multiscale approaches. Recently, Wouwer [9] has demonstrated the validity of this generalised approach to texture analysis.

1.3 Structure of the Dissertation

The body of this dissertation is organised as follows. The first two chapters present some background relating to CAD for digital mammography, as well as a description of the problem and the hypothesis. Chapter One presents the problem description and proposed solution. Chapter Two introduces the reader to the "state of the art" in the classification of abnormal lesions in digitised mammograms, as well as CAD techniques in general. This discussion is specifically, but not exclusively, directed at the classification of microcalcifications in digitised mammograms.

The next three chapters present some background theory and discussion of the relevant signal processing and pattern recognition tools used in this work. A discussion on wavelet analysis is presented in Chapter Three. This provides some background relating to wavelets and the wavelet transform. Chapter Four presents a range of possible techniques and approaches to the analysis of image texture. Chapter Five presents a discussion on pattern recognition, including feature selection and optimisation, classical classification techniques as well as artificial neural networks.

The detailed discussion of the development of the classification system, including the methods and procedures used, is presented in Chapter Six.

The final chapters present the results, discussion and conclusions of the investigation. The results and corresponding discussion of the development and testing of the classification system are presented in Chapter Seven. Finally, Chapter Eight presents the conclusions as well as suggestions for future work.

Chapter 2

Literature Survey

Recently there has been extensive research into the development of computer-aided diagnostic tools for the analysis of mammographic images. Work relating to, among other subjects; direct digital mammography, the enhancement of mammograms, the detection of abnormal lesions and the classification of abnormal lesions has been conducted. A thorough summary of the state of the art is presented by Giger [10]. In this paper, CAD is defined as the diagnosis made by a radiologist who takes into consideration the results of a computerised analysis of radiographic images and uses the analysis as a "second opinion" in detecting lesions and in making diagnostic decisions. Giger presents the full scope of research relating to CAD for digital mammography, including: the enhancement of mammographic images, computerised detection methods, as applied to both mass lesions and microcalcifications, the classification of abnormal lesions, using both manual and computer extracted features, and the possibility of merging these techniques and tools into complete mammographic workstations. A workshop relating to digital mammography is also held every two years. The most recent of these was held in 1998 and the corresponding workshop proceedings are available from Kluwer Academic Publishers [11]. The proceedings from the previous workshop are also available from Elsevier Science [12]. Another source of valuable information relating to digital mammography is the Digital Mammography Homepage: *http://www.rose.brandeis.edu/users/mammo/digital.html.*

Due to the usefulness of CAD techniques for the classification of abnormal lesions in digitised mammograms, many computerised classification systems have been developed. The remainder of this chapter will discuss some of the more interesting approaches to the classification problem.

In many of the systems, researchers have taken advantage of the ability of radiologists to extract mammographic features from mammograms. These manually extracted features are subsequently merged by rule-based systems, discriminant analysis or neural networks, into a final determination of the likelihood of malignancy. Ackerman et al. [13] use 36 radiographic features analysed by a group of radiologists. The feature extraction system was in the form of a comprehensive questionnaire, completed by participating radiologists. The features were then classified into benign and malignant categories using Bayesian Analysis. Baker et al. [14] and [15] also use features derived from radiologist descriptions of lesion morphology based on the standardised BI-RADS (Breast Imaging Reporting and Data System). A total of 18 features were used, 10 of these were morphologic features extracted by the radiologist and 8 were features encompassing data from patient personal and family history. The morphological features include features specific to microcalcifications, features specific to masses, and generic features applying to both masses and microcalcifications.

In other systems, the characteristic shape of malignant mass lesions is used as a discriminatory feature for classifying mass lesions as benign or malignant. Kilday et al. [16] use a set of seven shape features extracted from breast lesions, as well as patient age to differentiate between three common breast lesions; fibroadenomas, cysts and cancers. The shape features are based on the properties of the tumour boundary which are quantified using a measure of radial length. Feature selection was accomplished using linear discriminant analysis. A Euclidean distance metric was used to determine group membership. Pohlman et al. [17] also use the shape of the tumour as a basis for classification via a set of morphological features extracted from masses in digitised mammograms. Hou et al. [18] use a spiculation-sensitive pattern recognition technique called radial edge gradient analysis. This technique quantifies the degree of spiculation of the boarder of the mass, and classifies mass lesions based on their degree of spiculation. In a similar treatment Giger et al. [19] investigate a system that differentiates between benign and malignant masses based on the degree of spiculation of the boundary of the tumour. An artificial neural network is then employed for the classification of the masses based on their degree of spiculation.

Shape analysis may also be employed for classification of clusters of microcalcifications as benign or malignant. Shen et al. [20] use shape analysis for the classification of microcalcifications. They develop a set of shape factors to measure the roughness of the contours of calcifications. Moments, Compactness and Fourier Descriptors are used as the shape features.

Manually extracted features are also frequently used for the classification of microcalcifications. Wu et al. [21] employ an artificial neural network to distinguish between benign and malignant abnormal lesions based on a set of 14 radiologist extracted features. The performance of this system, based on the area under the receiver operating characteristic (ROC) curve, was found to be better than the average performance of attending resident radiologists. This classification system is also incorporated into an "intelligent" mammographic workstation to assist radiologists in the diagnosis of breast cancer [22]. Jiang et al. [23] utilise a set of features automatically extracted from digitised mammograms, together with a three-layer artificial neural network for the classification of clusters of microcalcifications as benign or malignant. Aghdasi et al. [24] also extract a set of features directly from clusters of microcalcifications. They evaluate over 100 numerical features which quantify the size, shape, number, roughness and configuration of the clusters of microcalcifications. They also determine a reduced subset of the most effective features for discriminating between benign and malignant microcalcifications.

Many interesting classifiers have also been investigated which improve the classification accuracy of the systems. Patrick et al. [25] use a network of expert learning systems for the classification of microcalcifications. The system uses features extracted from individual microcalcifications, microcalcification clusters and a set of clinical features. Hou et al. [26] evaluate three different types of classifier using four features automatically extracted from segmented mass lesions. They examine a two-step rule-based classifier, an artificial neural network and a hybrid system. The hybrid system which combines the rule-based system together with the neural network, outperforms the other systems. Brzakovic et al. [27] develop a fuzzy pyramid linking system together with a classification hierarchy which utilises Bayesian techniques to identify benign and malignant tumours at each level of the analysis.

Researchers have also investigated the possibility of performing an analysis on the actual tissue or texture of the lesion visible in the radiographic view. This information is then used as a cue for classification. Burdett et al. [28] propose a fractal method, which involves computing the fractal dimension over the entire lesion. This approach is based on the observation that malignant lesions exhibit rougher intensity profiles than benign lesions, therefore the fractal dimension offers a natural tool to assist radiologists in the diagnosis of abnormal mass lesions. Texture analysis techniques are especially useful for the classification of microcalcifications. Dhawan et al. [7], define a set of image structure features for the classification of malignancy. Two categories of grey-level image structure features are defined; the first include second order histogram statistics, representing the global texture, and wavelet decomposition-based features representing the local texture of the microcalcification region of interest (ROI). The second category of features represent first-order grey-level histogram-based statistics, and the size, number and distance features of the segmented microcalcification region. The wavelet analysis was performed on rectangular regions containing the microcalcifications. Kocur et al. [8] also extract wavelet-based features from ROI's in digital mammograms, and classify the microcalcifications using an artificial neural network.

This last group of texture-based classification systems, which employ texture analysis techniques to distinguish between benign and malignant classes of microcalcifications exhibit interesting properties and a number of these techniques are investigated and compared with the proposed texture analysis procedures investigated in this study.

Chapter 3

Wavelets

3.1 Introduction

For many years the Fourier Transform has been the most useful and effective technique for the analysis of signals. However, recently a new method for the decomposition and analysis of signals has emerged, known as Wavelet Analysis. Wavelet theory is not really a "new theory", rather a result of some cross-fertilisation between different fields, like mathematics, signal processing and geophysics, and the generalisation of some well established concepts.

The term "wavelet" was first introduced by Morlet, for the analysis of seismic data, who later collaborated with Grossman. Together they developed a more rigorous mathematical basis for these ideas [29]. This triggered the start of the development of a complete mathematical framework, which is now known as wavelet theory. Later Ingrid Daubechies [30] introduced families of orthonormal wavelets with compact support. Stephane Mallat incorporated the wavelet transform into the framework of multiresolution signal decomposition, and provided a fast implementation of the transform based on filter theory [31]. These important developments opened up the field of wavelet analysis to a wide range of applications in many areas including; physics, signal and image processing, signal and image coding and compression and

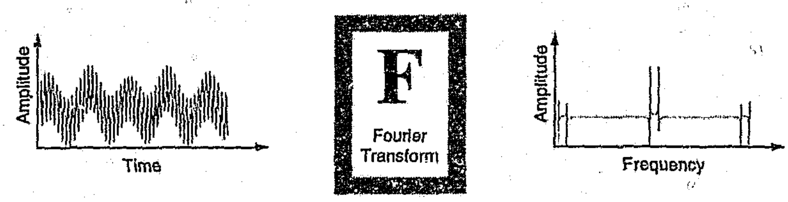

Figure 3.1: Diagrammatic representation of the Fourier Transform [35].

pattern recognition and segmentation.

A comprehensive mathematical description of wavelets and the wavelet transform is beyond the scope of this dissertation and the reader is referred to some of the many excellent books and papers available on the subject ([30], [31], [32], [33] and [34]). The remainder of this chapter will introduce the concept of wavelet analysis and multiresolution signal decomposition which provides a foundation for the application of these concepts in the remaining chapters.

3.2 Basic Outline of Wavelet Analysis

A convenient way of introducing wavelet analysis is as a development and generalisation of Fourier Analysis. In the Fourier Transform a signal is broken up into constituent sinusoidal components of different frequencies. In other words, it is a mathematical technique for transforming our view of the signal from a time-based one to a frequency based one. The Fourier Transform is diagrammatically represented in Fig. 3.1.

A drawback of the Fourier Transform is that during the transformation to the frequency domain, all time information is lost, i.e. it is impossible to tell when a particular event occurs by looking at the Fourier representation of a signal. If a signal is stationary, i.e. it doesn't change much over time, then this drawback is not that significant. However many interesting signals contain numerous non-stationary or transitory characteristics, and the Fourier transform is unsuitable for the analysis of these types of signals as it is incapable of detecting transitory and non-stationary behaviour.

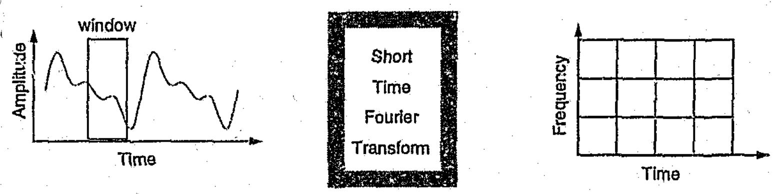

Figure 3.2: Diagrammatic representation of the STFT [35].

An adaptation of the Fourier Tansfrom, known as the Short-Time Fourier Transform (STFT) overcomes this deficiency by analysing only a small portion of the signal at a time. The STFT maps the signal into a two-dimensional representation in which both time and frequency information is available. The STFT represents a compromise between the time-based and frequency-based views of the signal. However the precision of the STFT is limited by the size of the window used for the analysis. Thus, the weakness of this approach is that once a window size has been selected, it is used for all frequencies. A diagrammatic representation of the STFT is depicted in Fig. 3.2.

The wavelet transform (WT) is the next logical step as it incorporates variable window sizes capable of analysing different frequencies. Longer time intervals are used for the precise analysis of low frequency information and short time intervals provide precise high frequency information. Thus, although it is not possible to simultaneously obtain arbitrarily fine localisation in both time and frequency due to the uncertainty principle, it is possible to study the slowly varying properties of the signal (low frequencies) over a longer time span and vice versa for the high frequency components of a signal. The difference between the STFT and the WT can be seen diagrammatically in Fig. 3.3 which picts the tiling of the time-frequency plane in each case.

A wavelet is essentially a limited duration signal that integrates to zero, and in the implementation of the wavelet transform it is used as the basis functions in much the same way sinusoidal functions are used as the basis functions in the computation of the Fourier Transform. The Fourier Transform is computed by integrating the signal to be analysed by a complex exponential. The results of this transform are the Fourier Coefficients, which when multiplied by a sinusoid of appropriate frequency

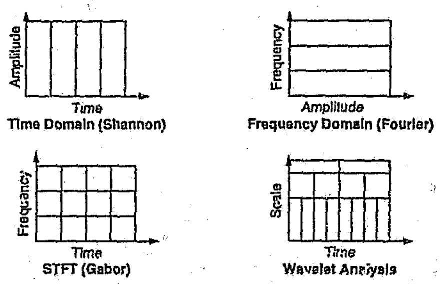

Figure 3.3: Tiling of the time-frequency domain for the STFT and the WT.

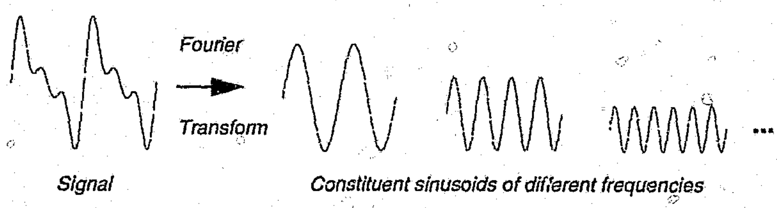

Figure 3.4: Graphical representation of a signal decomposed using the Fourier Transform.

yield the constituent sinusoidal components of the original signal. This is shown graphically in Fig. 3.4.

Similarly the wavelet transform is computed by integrating the signal to be analysed, multiplied by shifted and scaled versions of the wavelet function used, resulting in set of wavelet coefficients which are functions of scale and position. Thus, the wavelet transform provides both spatial and frequency information at different scales. Each of the wavelet coefficients may also be multiplied by the appropriately shifted and scaled wavelet to give the constituent wavelets of the original signal. This is shown graphically in Fig. 3.5.

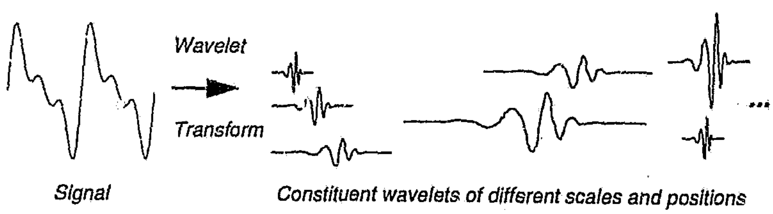

Figure 3.5: Graphical representation of a signal decomposed using the Wavelet Transform.

3.3 The Continuous Wavelet Transform

The continuous wavelet transform (CWT) of a 1-D signal $f(x)$ is defined as

$$C_{a,b} = \int f(x)\psi_{a,b}(x)dx \tag{3.1}$$

where the wavelet $\psi_{a,b}(x)$ is computed from the *mother wavelet* by translation and dilation, i.e.

$$\psi_{a,b}(x) = \frac{1}{\sqrt{a}}\psi\left(\frac{x-b}{a}\right) \tag{3.2}$$

Thus, a family of wavelets are used as the analysing functions in place of the complex sinusoids used in the computation of the Fourier Transform. The family of wavelet functions are derived from the *mother wavelet*, $\psi_{a,b}(x)$ which is indexed by the parameters a and b. The parameter a gives the dilation (which is inversely proportional to frequency or scale) and the parameter b gives the displacement (time localisation). The continuous wavelet transform is thus the sum over all time of the signal multiplied by scaled (dilated), shifted (displaced) versions of the mother wavelet. This process produces wavelet coefficients that are functions of scale and position. These coefficients constitute the results of a regression of the original signal performed on the wavelets.

3.4 The Discrete Wavelet Transform and Multiresolution Analysis

The CWT may be discretised by restraining a and b to a discrete lattice ($a = 2^n, b \in Z$). The mother wavelet then becomes

$$\psi_{n,m}(x) = 2^{-n/2}\psi(2^{-n}x - m) \tag{3.3}$$

To construct the mother wavelet $\psi(x)$ we may first determine a scaling function $\phi(x)$, which satisfies the two-scale difference equation

$$\phi(x) = \sqrt{2}\sum_k h(k)\phi(2x-k) \tag{3.4}$$

The corresponding mother wavelet $\psi(x)$ can then be constructed using the scaling function as follows

$$\psi(x) = \sqrt{2}\sum_k g(k)\phi(2x-k) \tag{3.5}$$

where

$$g(k) = (-1)^k h(1-k) \tag{3.6}$$

The coefficients $h(k)$ in (3.4) have to meet several conditions for the set of basis functions in (3.3) to be unique, orthonormal and have a certain degree of regularity. There are several sets of coefficients $h(k)$ in the literature that satisfy these conditions. Typically it is imposed that the transform be non-redundant, complete and that it constitutes a multiresolution representation of the original signal. Under these constraints an efficient real-space implementation of the transform using quadrature mirror filters exists [31]. Thus, the decomposition is viewed as passing the signal through a pair of filters H and G, with impulse responses $\widehat{h(n)}$ and $\widehat{g(n)}$, and downsampling the filtered signals by two. The impulse responses, $\widehat{h(n)}$and $\widehat{g(n)}$ are defined as

$$\widehat{h(n)} = h(-n),\ \widehat{g(n)} = g(-n) \tag{3.7}$$

The pair of filters H and G correspond to the halfband lowpass and highpass filters respectively, and are called quadrature mirror filters. This very practical filtering algorithm yields a *fast wavelet transform*, which is an efficient method of computing wavelet coefficients quickly and easily. A schematic diagram of the filtering procedure is shown if Fig. 3.6. In Fig. 3.6 the original signal S (a sinusoid with some high frequency noise) passes through the filter bank, and emerges as two signals which are then downsampled. The high frequency components or detail coefficients

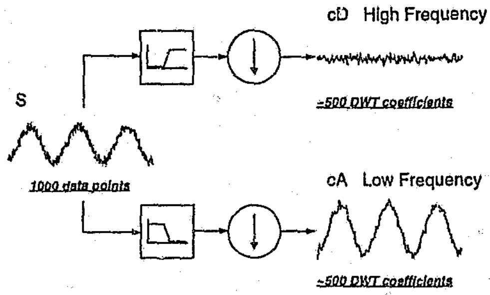

Figure 3.6: Fast WT implementation using a filter bank.

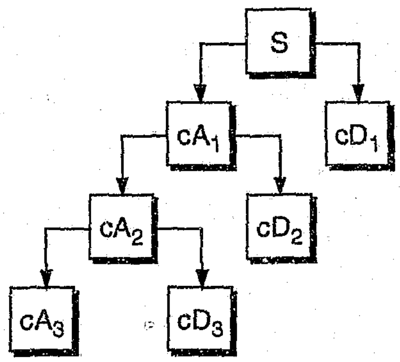

Figure 3.7: Multiple-level wavelet decomposition tree.

emerge from the high pass filter and the low frequency components or approximation coefficients emerge from the low pass filter.

The decomposition can also be iterated with successive approximations being decomposed in turn, which leads to a multi-level wavelet decomposition tree (Fig. 3.7). Where A and D are the approximation and detail coefficients respectively, at various levels of the decomposition. The signal under analysis is thus decomposed into the various components of the original signal at different scales or resolutions.

The extension to the 2-D case is usually performed by using a product of 1-D filters. In practice the transform is computed by applying a separable filter bank to the image as follows

$$L_n(\tilde{b}) = \left[H_x \left[H_y \star L_{n-1}\right]_{\downarrow 2,1}\right]_{\downarrow 1,2} (\tilde{b})$$

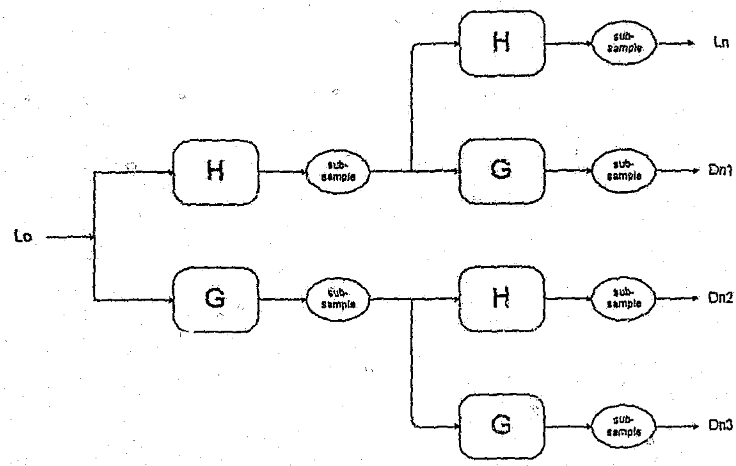

Figure 3.8: Separable filter bank implementation of the wavelet transform for a 2-D signal.

$$D_{n1}(\tilde{b}) = \left[H_x\left[G_y \star L_{n-1}\right]_{\downarrow 2,1}\right]_{\downarrow 1,2}(\tilde{b})$$

$$D_{n2}(\tilde{b}) = \left[G_x\left[H_y \star L_{n-1}\right]_{\downarrow 2,1}\right]_{\downarrow 1,2}(\tilde{b})$$

$$D_{n3}(\tilde{b}) = \left[G_x\left[G_y \star L_{n-1}\right]_{\downarrow 2,1}\right]_{\downarrow 1,2}(\tilde{b}) \tag{3.8}$$

where $\star$ denotes convolution, $\downarrow 2,1$ denotes subsampling along rows and $\downarrow 1,2$ denotes subsampling along columns. H and G are the low and high pass filters respectively. L_n is obtained by lowpass filtering, and is therefore referred to as the lowpass residue at scale n, or the approximation at scale n. The D_{ni} subimages are obtained by high-pass filtering in a specific direction and thus contain directional detail information at scale n, and are referred to as the detail images, providing horizontal, vertical and diagonal details. The original image is thus represented by a set of subimages at several scales; $\{L_d, D_{ni} | i = 1,2,3, n = 1..d\}$ which is a multiresolution representation of depth d of the original image. This separable filter bank for implementing the wavelet transform on a 2-D signal is shown in Fig. 3.8.

3.5 Concluding Remarks

The wavelet transform is an effective tool for signal processing, and in this instance is particularly useful for performing multiresolution signal decomposition. Furthermore, a multiscale analysis of both signals and images can be performed using a straightforward and efficient filter bank implementation of the transform. Therefore wavelets were selected as an ideal tool for performing multiscale texture analysis, which will be discussed in detail in the following chapter.

Chapter 4

Texture Analysis

4.1 Introduction

Texture is an important characteristic for the analysis of many types of images, including natural scenes, remotely sensed data and biomedical modalities. In addition, the perception of texture is believed to play an important role in the human visual system for recognition and interpretation. Therefore, the analysis of texture is of fundamental importance to image processing, computer vision and pattern recognition applications. For example: industrial monitoring of product quality, remote sensing of natural resources and medical diagnosis with computer tomography [36]. Consequently, a wide variety of approaches are available for the analysis of image texture. In this chapter an introduction to the various methods and approaches available for texture analysis will be presented. The proposed multiscale statistical approach to texture analysis will also be discussed.

4.2 Definition of Texture

Many papers that discuss texture analysis begin with a statement that there is no formal definition of texture. Others, give a definition that is suited to their

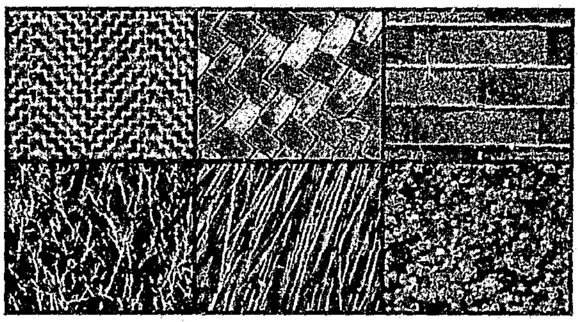

Figure 4.1: Some examples of Brodatz textures.

particular approach to texture analysis. Yet, there is no single, unambiguous and widely accepted definition of texture. This is as a result of the largely intuitive nature of the concept of texture, which is difficult to capture in a formal definition. Intuitively, the term "texture" is used to point to the intrinsic properties of surfaces, especially those that don't have smoothly varying intensity. It includes properties like smoothness, roughness, granulation and regularity. Some examples of different types of textures available as part of the Brodatz Album are shown in Fig. 4.1.

More formally, texture may be defined as an attribute representing the spatial arrangement of the grey levels of the pixels in a region [37]. Another definition is that the term texture generally refers to repetition of basic elements called texels. The texel contains several pixels, whose placement could be periodic, quasi-periodic or random. Natural textures are generally random, whereas artificial textures are often deterministic or periodic [38]. Texture may therefore be characterised by the spatial distribution of the grey levels in a particular neighbourhood in an image. It should also be noted that texture is not defined for a single point. Furthermore, the resolution at which an image is observed determines the scale at which the texture is perceived. For our purposes we can define texture as a repeating pattern of local variations in image intensity which is too fine to be distinguished as separate entities at the resolution at which it is observed.

4.3 Methods of Texture Analysis

There are three primary issues relating to texture analysis: texture classification, texture segmentation and shape recovery from the texture in an image. In this investigation texture analysis techniques are used for the classification of texture. In texture classification, the aim is to identify a given texture class from a range of texture classes. For example the task of recognising a particular region in an aerial image as belonging to either an agricultural, forest or urban region. Each region has a unique texture characteristic, and texture analysis algorithms are used to extract distinguishing features from the region to determine to which texture class it belongs. Statistical methods are extensively used for texture classification, and properties such as grey-level co-occurrence, contrast, homogeneity and entropy are regularly computed to facilitate texture classification.

Of the three popular existing approaches to texture classification, namely statistical techniques, structural techniques and model-based techniques, the statistical methods are the most widely used. These statistical methods are utilised in this investigation, and will be introduced in the next section. Briefly, structural techniques describe the texture using texture primitives and syntactic rules. Model-based approaches determine an analytical model of the texture, for example, using Markov random fields. For reviews on existing texture analysis methods see [39].

4.3.1 Statistical Techniques

Since texture is a spatial property, a simple one-dimensional histogram, or first-order grey-level histogram, is not useful in characterising texture. For example, an image in which pixels alternate from black to white in a checkerboard fashion will have the same histogram as an image in which the top half is black and the bottom half is white. The perception of image texture is related to the spatial (statistical) distribution of the grey tones in an image. In order to capture this grey-level spatial dependence, a two dimensional dependence matrix know as a co-occurrence matrix can be computed. The co-occurrence matrix is capable of quantifying the spatial dependence of the grey-levels in an image. These probability-distribution matrices thus describe the statistical nature of the texture in the spatial domain [40].

Another measure that is useful for quantifying the statistical nature of the texture in an image is the autocorrelation function. A brief description of these techniques follows, for a comprehensive discussion of the grey-level co-occurrence matrix see [6].

Grey-Level Co-Occurrence Matrices

The first-order grey-level histogram is defined as the distribution of the probability of occurrence of grey-levels in an image. The second-order histogram, or grey-level co-occurrence matrix, $H(y_i, y_j, d, \theta)$ represents the distribution of probability of occurrence of a pair of grey-level values separated by a given displacement vector, d, at an angle of θ. In other words, $H(y_i, y_j, d, \theta)$ indicates the frequency with which particular grey-level pairs y_i and y_j, at an angle of θ, are separated by the displacement vector d. This is computed for all grey-level pairs in the image, for various displacement vectors and for all angles (usually 0°, 45°, 90° and 135°, around a central pixel). Four co-occurrence matrices are thus calculated for each displacement vector used. Typically these angularly dependent matrices are not used directly, rather the mean of the four co-occurrence matrices are used, ensuring rotational invariance [6]. The elements of $H(y_i, y_j, d, \theta)$ are normalised by dividing each entry by the total number of pixel pairs and thus enabling the co-occurrence matrix to be treated as a probability mass function.

Autocorrelation

The autocorrelation function $p[k, l]$ for an $N \times N$ image is defined as:

$$p[k,l] = \frac{\frac{1}{(N-k)(N-l)} \sum_{i=1}^{(N-k)} \sum_{j=1}^{(N-l)} f[i,j]\, f(i+k, j+l)}{\frac{1}{N^2} \sum_{j=1}^{N} \sum_{j=1}^{N} f^2[i,j]} \quad where\ 0 \le k,l \le N-1 \tag{4.1}$$

For images comprising repetitive texture patterns the autocorrelation function exhibits periodic behaviour, with a period equal to the spacing between adjacent texture primitives. When the texture is coarse, the autocorrelation function drops off slowly, and when the texture is fine, it drops off more rapidly. The autocorrelation function is useful as a measure of the periodicity of texture and as a measure of the scale of the texture primitives in an image.

4.3.2 Multiscale Techniques

Statistical texture analysis techniques exploit the local correlation between image pixels as a method of classifying the texture in an image. However these techniques are implemented at a fixed scale, and an important property of image texture is the scale at which the texture is analysed. The validity of a multiscale approach to texture analysis has been verified by studies which show the human visual system processes images in a multiscale fashion. Therefore multiresolution techniques, which transform images into a representation in which both spatial and frequency information is present at different scales, are important tools for the analysis of image texture [41] and [42].

A number of systems have been developed that employ multiscale techniques for the analysis of image texture. In particular Gabor filters have been employed for texture segmentation. For a review of these techniques see [9], [41] and [42]. However, recently wavelet theory has emerged as a concise mathematical framework for multiscale analysis. Wavelet techniques can be used to transform images into a representation in which both spatial and frequency information is available, and therefore wavelet analysis has emerged as a powerful signal and image processing tool. For texture analysis applications, the wavelet transform is used to map the image into a series of detail and approximation images at different resolutions. This multiresolution analysis is analogous to passing the signal through a multichannel filter. Textural information, in the form of the energy or entropy extracted from these various resolution sub-images, can be used for classification or segmentation [9], [36] and [42]. M-band wavelet decomposition, a direct generalisation of the standard 2-band wavelet decomposition has also been applied to the discrimination of natural textures [43].

Multiscale texture analysis techniques have also been applied to the detection and classification of microcalcifications in digitised mammograms. Strickland [44] presents a wavelet-based approach for the detection and segmentation of microcalcifications. More recently these multiscale texture classification techniques have also been applied to the classification of microcalcifications in digitised mammograms [7] and [8].

4.3.3 Multiscale Statistical Techniques

In this investigation, a third approach to texture analysis is also proposed. This technique is based on a combination of the statistical and multiscale approaches to texture analysis. The hypothesis is that texture can be more completely characterised based on a statistical analysis of its multiscale representation. By combining the statistical and multiscale approaches, it is possible to analyse the local correlation between pixel values at various different scales or resolutions.

The multiscale statistical texture analysis is performed by first using the wavelet transform to decompose the image into a number of different resolution sub-images. A statistical analysis of the texture contained in these sub-images is then performed. The co-occurrence matrices are used to perform the statistical analysis.

This approach has also recently been tested in a classification problem involving natural textures [9].

4.4 Concluding Remarks

In this investigation a variety of these statistical and multiscale techniques are used for the classification of the texture surrounding microcalcifications in digitised mammograms. The suitability of this texture-based approach to the classification of microcalcifications will emerge during the course of the investigation. In addition, the most suitable texture analysis technique for this classification problem will also come to light.

Chapter 5

Pattern Recognition

5.1 Introduction

Pattern recognition may be broadly defined as the biological and mental processes that enable external signals stimulating our sense organs to be converted into meaningful perceptual experiences, or somewhat more restrictively, how we are able to assign a name of some kind to a complex stimulus. This process occurs in the following sequence. Firstly, the pattern must be perceived by our sense organs. Secondly, in order for this pattern to represent a meaningful perceptual experience, a similar pattern, or a pattern in the same class, must have been previously perceived. Finally, the past perception of the pattern must be remembered, and some correspondence drawn between the past and the present perception [45]. A concise definition is also given by Vapnick [46]:

> "A person (the instructor) observes occurring situations and determines to which of c classes each one of them belongs. It is required to construct a device (the classifier) which, after observing the instruction procedure, will carry out the classification approximately in the same manner as the instructor."

The patterns, or "occurring situations", to be classified may be any pattern that it is possible for a human to recognise, however, the field of pattern recognition is not restricted to objects or information that can be recognised by humans. Any information or abstract object that may be represented by an array of numbers, e.g. a sampled waveform or a scanned image can be used as a feature vector, provided that the feature vector provides sufficient discriminatory information to distinguish between the classes in the particular problem.

An excellent example of a set of measurements describing particular classes in a pattern classification problem is the "Iris" data set. In 1935 Anderson collected measurements of the petal and sepal leaves of three different species of the Iris family Iridaceae; setosa, versicolor and virginica. This data was later used by Fisher to illustrate the concept of Linear Discriminant Functions. Four different features were investigated: sepal length, sepal width, petal length and petal width. For each of the three flower classes 50 samples were collected, totalling 150 samples. Since then this data set has turned into a classic problem in pattern recognition.

The real world objects in this example are species of flowers, i.e. there are three distinct classes. The sensor in the example is simply a device which measures length. And the feature vector is simply the four measurements. The classification task is therefore to distinguish between the different flower groups based on the information contained in the four measurements taken of each of the sample flowers. The Iris dataset is available via ftp from *ftp://ftp.ics.uci.edu/pub/machine-learning-databases/*.

There are two types of pattern recognition problems; *supervised learning* problems, and *unsupervised learning* problems. In pattern classification systems based on supervised learning, the class label of each sample is known, and this information is used in the design and implementation of the classification system . In the unsupervised learning case the class labels of each sample are unknown, and the classes must be inferred from the available data. Unsupervised learning problems are dealt with using segmentation algorithms and will not be discussed here.

This chapter presents an introduction to pattern recognition, as well as a discussion on some important aspects and techniques available for pattern recognition. The reader is also referred to some of the many good textbooks on the subject, namely: [45], [47], [48] and [49].

5.2 The Bayes Classifier

The Bayes rule for decision making is the theoretical optimum performance that a classifier can achieve. Practical implementations of classifiers attempt to reach the performance of the Bayesian classifier. The most relevant quantity for the classification of a feature vector $\vec{x}$ is the probability that it is derived from an object belonging to class w_i. This can be calculated from the likelihoods and priors using *Bayes formula* which is defined as follows:

$$P(w_i \mid \vec{x}) = \frac{P(w_i)p(\vec{x} \mid w_i)}{p(\vec{x})} \tag{5.1}$$

where, $P(w_i \mid \vec{x})$ is the *a posteriori* probability, i.e. the probability that given a vector $\vec{x}$, it belongs to class w_i, $P(w_i)$ is the *a priori* probability of occurrence of an event from class w, $p(\vec{x} \mid w_i)$ is the class conditional probability density function and $p(\vec{x})$ is the unconditional probability density function. Bayes formula defines a method for calculating the *a posteriori* probability that a given pattern belongs to a certain class.

The decision rule of the classifier is the function $\hat{w}(\vec{x}) = w_i$, i.e. the function that assigns a class label to each feature vector.

A loss function L_{ij} is also required. This is defined as some arbitrary cost associated with a misclassification. Frequently, the function $L_{ij} = 1 - \delta_{ij}$ is used, thus a loss of *0* is assigned to a correct classification and a loss of *1* is assigned to a misclassification. (Where, δ_{ij} is the *Kronecker delta*: $\delta_{ij} = 1$ if $i = j$ and 0 otherwise.)

The conditional risk is defined as the expected loss incurred by assigning the sample vector $\vec{x}$ to class w_j and may be represented as:

$$r_{w_j}(\vec{x}) = \sum_{i=1}^{c} L_{ij} P(w_i \mid \vec{x}) \tag{5.2}$$

The Bayes Classifier therefore decides which class to assign a given sample based on the decision rule given by:

$$\widehat{w}(\vec{x}) = w_i \ if \ r_{wi}(\vec{x}) \leq r_{wj}(\vec{x}) \tag{5.3}$$

If we use the loss function given above this gives an equivalent rule given in terms of the *a posteriori* probabilities:

$$\widehat{w}(\vec{x}) = w_i \ if \ P(w_i \mid \vec{x}) \geq P(w_j \mid \vec{x}) \ \forall j \tag{5.4}$$

This may also be viewed in terms of the conditional probability density functions, or *class likelihoods*, by substitution into Bayes Formula (5.1). It is possible to design the Bayes Classifier if either the conditional likelihoods or the *a posteriori* probability densities are known.

5.3 Parametric Classifiers

Parametric classifiers assume that the class conditional likelihoods have a known functional form which depends on a few parameters. The classifier is designed by estimating the distribution parameters from the training set. New data is then classified using the decision rule defined in (5.3). It is thus possible to design many different classifiers based on different forms or combinations of different parametric distributions. A popular parametric classifier is based on the Gaussian distribution. This type of classifier is recommended for cases when the data obeys a specific distribution.

5.4 Non-Parametric Classifiers

In contrast with parametric classifiers non-parametric classifiers do not assume an *a priori* known parametric form for the class likelihoods $p(\overrightarrow{x} \mid w_i)$, but estimate them directly from the design samples. Note that a parametric classifier is designed by estimation of the required density parameters. After design, all relevant densities are known in feature space and a sample is classified using the Bayes decision rule. Non-parametric classifiers operate by estimating the density around that particular sample directly from the design set, and classifying the sample accordingly. Thus non-parametric classifiers implement the decision rule locally and the likelihoods need to be estimated for each sample offered to the classifier.

This procedure can be explained as follows. Suppose we have N design samples and need to estimate the density around a feature vector $\overrightarrow{x}$ in feature space. This can be estimated by constructing a region S around $\overrightarrow{x}$ and then counting the number of design samples K that fall within this region, and denoting V for the volume of S. This leads to the following estimate for the density $p(\overrightarrow{x})$:

$$p(\overrightarrow{x}) \simeq \frac{K}{NV} \tag{5.5}$$

This estimate depends on N and the region S. The only variable available is the size of the region S. The larger the region, the more samples it will contain, and the more accurate the estimate will be [9].

Given a design set and a sample to be classified, there are two approaches to evaluate (5.5). We can either fix the region S and count the number of samples falling within S, this is know as the *Parzen window* approach. Or fix K and determine the volume V, which is know as the *k-nearest neighbour* approach. The k-nn classifier is used later as part of the classification scheme and will now be introduced.

5.4.1 K-Nearest Neighbour Classifier

The nearest neighbour classifier exchanges the need for a knowledge of the underlying distributions for the knowledge of a large number of correctly classified patterns [45]. The basic idea behind the nearest neighbour classifier is that samples which lie close together in feature space are likely to either belong to the same class or to have about the same *a posteriori* distributions of their respective classes.

To estimate the distribution or density around a feature vector, its K nearest neighbours from the design or training set are sought. The distance between the sample feature vector and each of the nearest neighbours is then calculated. By doing this one assumes that the feature space is a metric space; i.e. there exists some function $d(\vec{x}, \vec{y})$ which expresses the distance between two points $\vec{x}$ and $\vec{y}$ in feature space. Frequently the *Mahalanobis distance* is used as a distance metric, but the *Euclidean distance*, a specialised version of the Mahalanobis distance is also frequently used.

Once again, if there are N samples belonging to c classes with N_i samples in class w_i, then to classify a feature vector $\vec{x}$, we find the K samples closest to it and determine the volume V in which these samples reside. And suppose that there are K_i samples of class w_i among the K nearest neighbours, then the class likelihoods, the unconditional probability density and prior probability can be estimated as follows:

$$p(\vec{x}|w_i) \simeq \frac{K_i}{N_i V} \quad p(\vec{x}) \simeq \frac{K}{NV} \quad p(w_i) \simeq \frac{N_i}{N} \tag{5.6}$$

If these estimates are used together with the Bayes decision rule (5.4), rewritten in terms of the class likelihoods, the following classification rule results:

$$\widehat{w}(\vec{x}) = w_i \quad if \;\; p(\vec{x}|w_i)p(w_i) > p(\vec{x}|w_j)p(w_j) \quad \forall j$$

$$\Longrightarrow K_i > K_j \quad \forall j \tag{5.7}$$

Therefore, this leads to a very simple classification procedure. The sample $\vec{x}$ is assigned to that class to which most of its K nearest neighbours belong.

5.5 Neural Networks

An artificial neural network is a parallel, distributed information processing structure consisting of processing elements (which posses a local memory and carry out localised information processing operations) interconnected via unidirectional signal channels called connections. Each processing element has a single output connection that branches ("fans out") into as many collateral connections as desired; each carries the same signal, namely the processing element output signal. The processing element output signal can be of any mathematical type desired. The information processing that goes on within each processing element can be defined arbitrarily with the restriction that it must be completely local; that is, it must depend only on the current values of the input signals arriving at the processing element via impinging connections and on values stored in the processing element's local memory [52].

This section will present a basic overview of neural networks. Further discussion on the theoretical foundation, design and training of neural networks is available in [52] and [53].

A typical example of a neural network is the feedforward multilayer perceptron depicted in Fig. 5.1. It consists of several neurons grouped together in several layers. Each of the neurons is composed of a transfer function f and a summing node (see Fig. 5.2). Each neuron also has an associated weight w which is multiplied by the input vector p to the neuron. Provision is also made for a bias b. In operation the input vector p is multiplied by the weight w, which is added to the bias b and passed through the transfer function f yielding the neuron output a. The transfer function is usually a non-linear function such as a sigmoidal function. An example of a single neuron is shown in Fig. 5.2. When interconnected in a network, a different weight

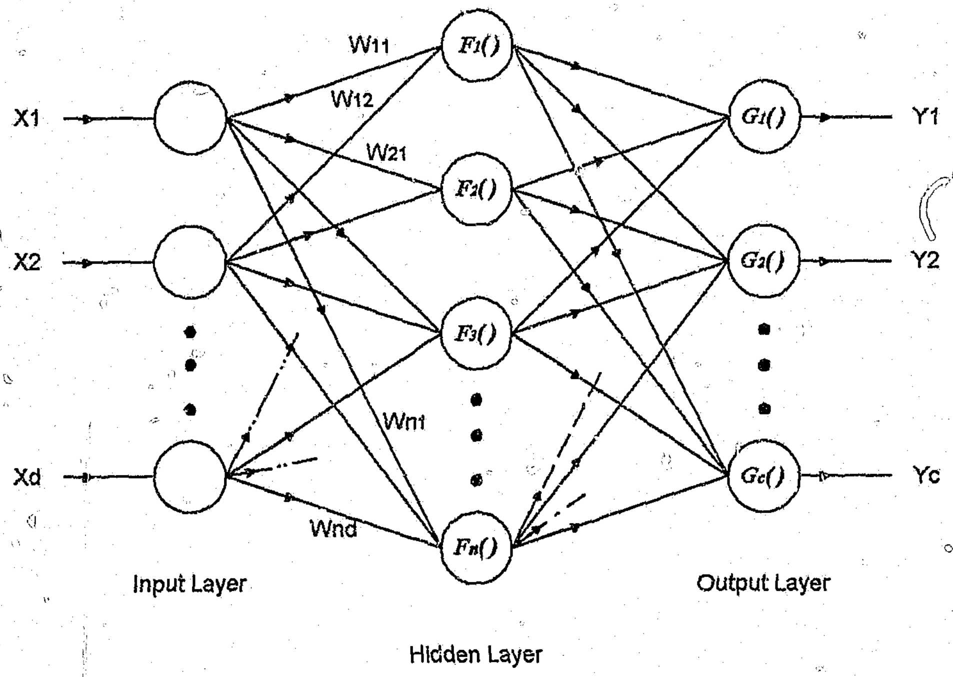

Figure 5.1: Example of a multilayer perceptron with three layers.

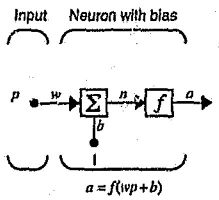

Figure 5.2: An example of a single neuron.

is associated with each connection between each neuron as shown in Fig. 5.1.

The number of layers, number of neurons in each layer and their interconnectivity is referred to as the network architecture. In this network the input feature vectors are received at the input layer, which acts merely as a buffering zone, and propagated through the hidden layer to the output layer. Each neuron processes the signal it receives and sends its output to the following layer.

The network is designed or trained as follows. Each of the weights is initialised with a normally random initial value. Input vectors from the training set, which have a known output, are fed into the network. The weights are then adjusted using the desired known outputs in such a way so that when a similar future input vector is fed into the network the output of the network will more accurately correspond to the desired output. This is usually performed by iteratively feeding the samples of the training set into the network and observing the output. The difference between this output and the desired output (i.e. the true class of the design sample) is used to compute an error criterion. The weights are then adjusted to minimise this criterion. The rule according to which the weights are adjusted is called the learning rule or training method. A popular learning rule is the *backpropagation learning rule*, which propagates the error backwards through the network in order to update the weights of the network.

Thus, a neural network partitions the feature space into several decision regions. And each sample feature vector is assigned to a class based on the current configuration of the network, which has been trained using known samples. The possible form these decision boundaries can assume is specified by the network architecture. The exact form of the decision boundary is determined by the weights which are adapted during training.

The reason for calling this classification scheme a "neural network" is that it bears some resemblance to the structure of the brain. In a simplified form the brain is composed of neurons interconnected via dendrites through which signals in the form of electro-chemical impulses propagate. Consequently neural networks have acquired a reputation for being able to mimic human intelligence. However, they simply represent a complex function dependent on their weights and transfer functions, which will assign each point in feature space to a particular class, i.e. the network is a model for the decision boundary [9].

5.6 Dimensionality Reduction

In this treatment of the field we have assumed that patterns are represented by d-dimensional measurement vectors obtained through some data acquisition mechanism. The choice of these features or measurements is however crucial to the success of the pattern classification system. Good features enhance within class pattern similarity and between class pattern dissimilarity. However, the more complex the patterns are, the more difficult it is to decide what constitutes a good measurement.

Although quantity can never compensate for quality, a method for discovering the best set of features to distinguish between a group of classes, is to utilise all possible information about the problem, by selecting all possible measurements that could possibly have some significance. However an increase in the number of features not only leads to a more complex classifier structure, but also the presence of possible redundant or irrelevant information, which could detrimentally affect the reliability of the classifier. In addition it should also be noted that given a large enough feature space relative to the number of samples in the classification problem, it is possible to artificially distinguish between the classes.

Thus, feature selection and extraction algorithms are implemented to quantify the amount of classification information provided by each feature, and select only those features that improve the performance of the classification system. It is also useful to quantify the amount of discriminatory information provided by groups of measurements, to reveal redundancy in the feature space.

Feature extraction algorithms work by combining the existing features and transforming the original feature space into a new lower dimensional optimised feature space. Two examples of feature extraction routines are Principle Component analysis and Linear Discriminant analysis. Details of these routines are available in [45].

In contrast, feature selection algorithms work by selecting only a few optimum features from the original feature space. A feature selection algorithm consists of two parts, viz. the selection criterion and the search strategy. The selection criterion, or "cost function", is the method of rating the ability of a feature to discriminate between the classes. The search strategy is the method of determining the optimum combination of the rated features.

Typically interclass or probabilistic distance measures are used as the selection criterion. The interclass distance measure works by computing the average interclass distance between all data points in feature space. The most successful features are obviously the ones which maximise the inter class distance and minimise the intra-class distance. The probabilistic distance measures are based on a measure of the similarity of the class conditional probability density functions. The "further" apart the density functions are the better the feature.

The optimum combination of the features rated using the selection criterion is determined using an algorithmic search strategy. An exhaustive search, i.e. for the optimum combination of the subset of features d, out of the total feature space D, requires $\binom{D}{d}$ criterion evaluations. This quickly leads to an excessive number of potential trials. Therefore, sub-optimum search strategies are often employed. A Unidimensional Search ("Best Features") selects the best d features from the total feature set, based on the rating by the selection criterion. A drawback of this algorithm is its blindness to correlated features.The sequential forward selection (SFS) starts with the best feature, and then tests to see which other feature, excluding the already selected feature scores the highest multivariate criterion. This feature is then included in the reduced feature space. This continues until the required number of features have been selected. The Sequential Backward Selection (SBS) works in a similar way to the previous algorithm, however it starts with the complete feature set and successively removes the least discriminatory feature, until the required number of features have been selected [45]. A recent review paper investigated the effectiveness of various different feature selection algorithms and the reader is referred to this paper for further details on the subject [54].

In this way, using a specific selection criterion together with one of the search strategies a reduced optimum feature space may be selected from the original feature space.

5.7 Concluding Remarks

In conclusion however, it should be borne in mind that a classifier is only as successful as the quality of the feature set. In other words, if a good feature set is available then any classifier will do. If not, then even the most sophisticated classifiers will fail to solve the pattern recognition problem.

In this investigation a number of different types of texture features are used to classify the regions surrounding microcalcifications in digitised mammograms as benign or malignant. A wide range of different types of texture features are extracted from each of the regions. This high dimensional feature space is then reduced using a feature selection algorithm. And the discriminatory ability of the features is tested using two different types of classifier, a k-nn classifier and a feedforward artificial neural network.

It should also be noted that the aim of this investigation is to demonstrate and test the ability of texture analysis techniques for the discrimination of benign from malignant microcalcifications. Thus any of the available classification systems and feature selection or extraction techniques could have been used, and the chosen system, described below is not the only possible configuration of the system.

Chapter 6

The Classification System

6.1 Introduction and Overview

The aim of the classification system is to distinguish between benign and malignant microcalcifications based on the radiographic view, i.e. to determine if it is possible to predict the outcome of a surgical biopsy based on an analysis of the mammographic film.

For the system three different sets of textural features are compared as to their suitability for the task of discriminating between the two microcalcification classes. Texture methods are investigated as they present a useful and effective method of tackling the problem as discussed above (see 1.2). The texture features are based on statistical and multiscale texture analysis techniques. The advantage of texture methods is that the discriminatory features may be extracted directly from the region of interest (ROI) containing the microcalcification, or cluster of microcalcifications. Therefore, the input to the classification system is simply a ROI containing microcalcifications. The region is either detected by an automatic detection scheme which may be coupled to this classification scheme, or manually highlighted by an attending radiologist. An example of a full mammographic image, along with a corresponding highlighted area containing a cluster of microcalcifications is shown in Appendix A.

The three groups of texture features are then extracted directly from the ROI. Firstly, a set of statistical texture features, based on Haralick's co-occurrence analysis. Secondly, three different types of wavelet texture signatures, as suggested by Laine [36] and Kocur [8]. Finally, a proposed third set of features, a set of statistical texture features extracted from a multiscale representation of the original ROI. The first set of features represent a statistical analysis of the texture in the ROI, while the second set represent a variety of wavelet-based multiscale texture signatures. The third proposed set of features represent a statistical analysis of the multiscale representation of the ROI. The extracted feature space is then used as input to either a k-nearest neighbour classifier or an artificial neural network which discriminate between benign and malignant classes of microcalcification. A schematic block diagram of the system as described is shown in Fig. 6.1.

In this chapter a detailed description of the classification system is presented. This includes a full discussion of the procedures used to construct each component of the system. The implementation of this classification scheme, and associated experimental results and discussion are presented in the following chapter.

6.2 Database Selection, Digitisation and ROI extraction

One of the difficulties associated with the comparison of different classification schemes is due to the inconsistent use of a standardised mammographic dataset. For this work two publicly available standardised mammographic databases were used. The Nijmegen Digital Mammogram Database [56], and the Lawrence Livermore National Laboratories / University of California at San Francisco (LLNL/UCSF) [57] database were used for training and testing. Both databases are widely used in the literature and are available for research purposes from the compilers thereof.

The mammograms in the Nijmegen Database consist of images from 21 different patients. For each case both cranio-caudal and mediolateral views are available. All

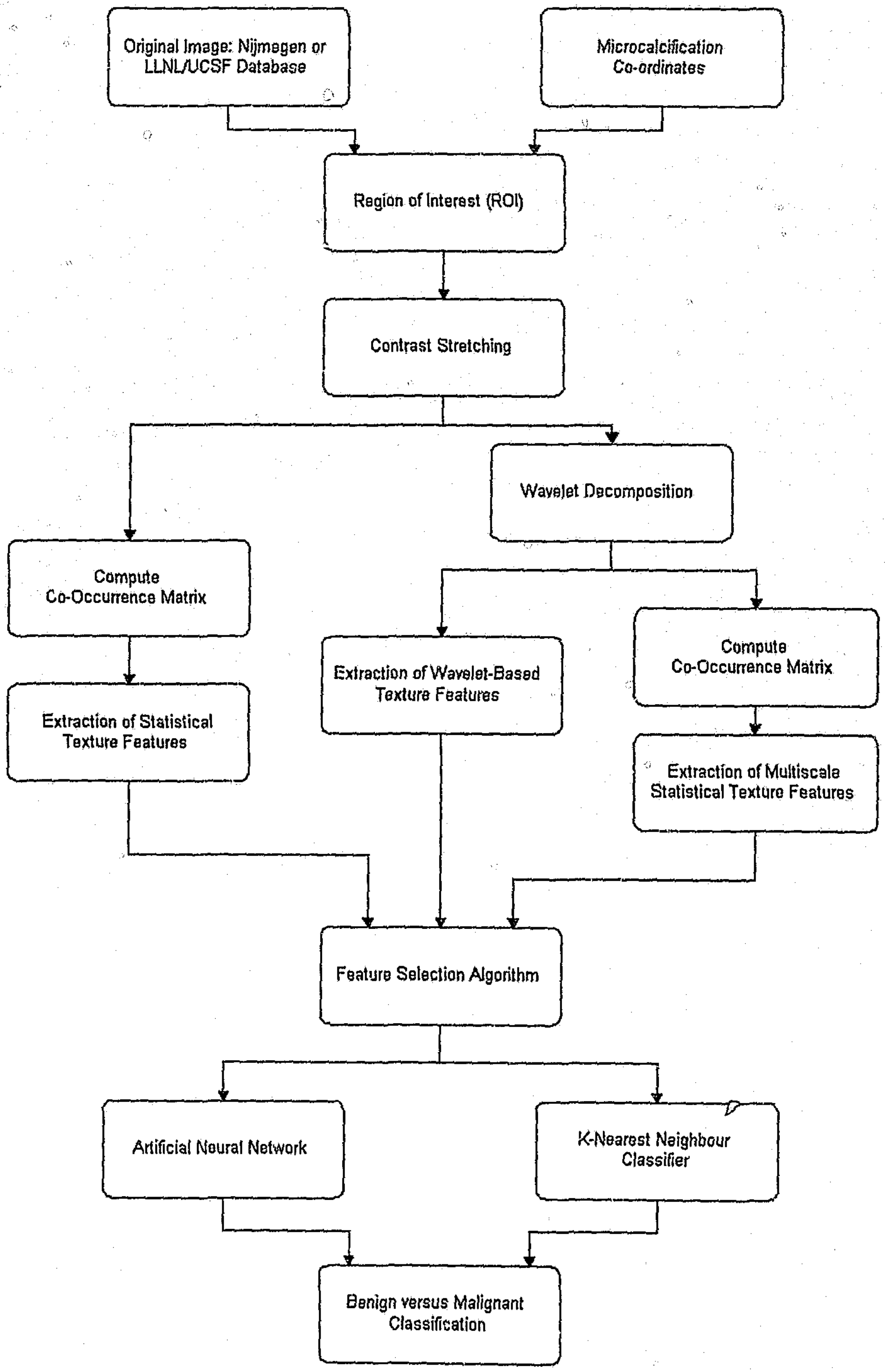

Figure 6.1: Schematic block diagram of the classification system.

views contain one or more microcalcification clusters. In total there are 104 microcalcifications or microcalcification clusters in the database. Truth information based on biopsy results is also available for each microcalcification cluster in the database. The images were digitised at a 100 microns/pixel resolution with 12 bits of pixel grey-level information. The position and size of the microcalcification clusters in each image were marked by two expert radiologists, based on all patient data available. These annotations are supplied in a separate file, containing the co-ordinates and radius of each microcalcification or microcalcification cluster. Using these annotations, appropriate regions of interest encircling each microcalcification could be extracted directly from the original mammograms. The size of the ROI's range from a minimum of 128 $\times$ 128 pixels to a maximum of 500 $\times$ 500 pixels. The contrast in the extracted ROI containing the microcalcification was stretched to a normalised grey-level range 0–255. The various texture features are then extracted from this normalised ROI. Of the 104 available microcalcification clusters 97 appropriate images were chosen for the analysis, comprising 29 benign clusters and 68 malignant clusters.

The LLNL/UCSF Database contains 198 films from 50 patients (4 views per patient, but only 2 views for one mastectomy case). The films were digitised at 35 microns/pixel, with 12 bits of grey-scale information. The films were selected to span a range of cases of interest, including 5 normal healthy cases, 5 normal but difficult cases, 20 cases of obviously benign microcalcifications, 12 cases of suspicious benign microcalcifications and 8 cases with malignant clusters of microcalcifications. Truth information is provided along with each image in the form of full size binary images showing the locations of the microcalcification clusters, and the lesion type and diagnostic information. Truth information was based either on a biopsy result, or on 3 years of follow up without change. Of the approximately 130 available microcalcification clusters in the database, 83 ROI's containing microcalcification clusters were extracted. The ROI's were manually selected from the original mammograms using the associated truth information. Of the 83 ROI's, 56 are benign and 27 are malignant. The size of the ROI's varies from a minimum of 128 $\times$ 128 pixels to a maximum of 370 $\times$ 370 pixels. These ROI's were also normalised to a 0–255 grey-level range.

Two datasets were used to test the robustness of the texture features. The datasets were not combined, and the classification system was evaluated separately on each dataset, i.e. each dataset was independently used for training and testing purposes.

The vario[illegible]xture features under analysis are then extracted from each of the ROI's in each of the datasets.

For this classification system it would also be possible to scan mammographic files directly and use ROI's extracted from the digitised mammogram as input to the classification system.

6.3 Extraction of Statistical Texture Features

The first group of texture features extracted from the ROI's are the features based on a statistical analysis of the texture in the area surrounding the microcalcifications.

Second-order statistics are commonly employed for the analysis of image texture. And the two-dimensional spatial dependence matrix, or co-occurrence matrix is used here to calculate these second-order statistics. The perception of image texture is related to the spatial (statistical) distribution of the grey tones in an image, and the co-occurrence matrix is capable of quantifying the spatial dependence of the grey-levels in an image. These probability-distribution matrices thus describe the statistical nature of the texture in the spatial domain [6]. A description of the grey-level co-occurrence matrix is described above (see 4.3.1).

The co-occurrence matrices were computed for each the four angles surrounding the central pixel under analysis (0°, 45°, 90° and 135°) and then averaged to obtain a single matrix invariant to rotation for each image. The analysis was performed on all ROI's in the two databases used in the study. The co-occurrence matrices were also computed for various different displacement vectors, as suggested in [9].

	Texture Feature
1	Contrast
2	Angular Second Moment
3	Entropy
4	Inverse Difference Moment
5	Maximum Probability
6	Cluster Shade
7	Cluster Prominence
8	Information Measure of Correlation

Table 6.1: Co-occurrence texture features

Textural information was then extracted from these computed co-occurrence matrices. Haralick [6] suggests a set of co-occurrence textural features. These measures are computed from the co-occurrence matrix to describe the texture in an image. Some of these measures relate to specific textural characteristics of the image such as homogeneity, contrast, entropy, energy and the presence of organised structure within the image. Other measures characterise the complexity and nature of the grey-tone transitions which occur in the image. These eight statistical texture features were extracted from each of the co-occurrence matrices computed from each of the sample ROI's. The eight texture features are listed in Table. 6.1. Formal definitions of these features are available in Appendix B. This statistical texture analysis therefore results in a set of 8 texture features which characterise the textural information in each sample ROI.

6.4 Extraction of Wavelet-Based Texture Signatures

It has been widely demonstrated that multiresolution techniques are an important aspect of texture analysis. Wavelet theory provides a precise and unified framework for multiresolution analysis, and therefore may be used for the analysis of image texture [36], [42]. For the second group of texture features, three sets of wavelet based texture signatures as proposed by [36] and [8] are extracted from the sample ROI's. The effectiveness of these features, as analytical tools for the analysis of the texture surrounding microcalcifications, is then compared.

The discrete wavelet transform is used to map the regions of interest to a series of detail coefficients, which constitute a multiscale representation of the ROI. Daubechies wavelets are orthonormal, regular wavelets of compact support and are therefore suitable for the analysis of signals with finite support, particularly for image analysis. Daubechies wavelets D_6 and D_{20} provide a good combination of regular prototype wavelets with varying sizes to extract texture information with varying spatial frequency [36]. Biorthogonal spline wavelet filters and D_4 wavelets, as used by Kocur et al. [8] are also implemented.

To obtain features that reflect scale-dependant properties, a separate feature is extracted from each scale (level), of the wavelet transform. The first texture signature computed is the energy [36]

$$energy = \frac{norm}{length \times breadth} \tag{6.1}$$

where, $norm = \sum_i \sum_j x_{ij}^2$, and x_{ij} are the computed coefficients of the wavelet transform (i^{th}row and j^{th}column). This provides a set of features consisting of energies of different scales, which is important for texture analysis [42].

A measure of the entropy of the decomposed wavelet coefficients was also computed [7]

$$entropy = -\sum_i \sum_j \left[\frac{x_{ij}^2}{norm}\right] \log \left[\frac{x_{ij}^2}{norm}\right] \tag{6.2}$$

Kocur et al. [8] also use an additional feature, the square root of the norm,

$$r = \sqrt{norm} \tag{6.3}$$

A wavelet decomposition was performed on each of the ROI's in each of the databases. The wavelet texture signatures were computed for each of the approximation and detail coefficients at each level of the decomposition. The wavelet transform was computed to 4 levels of decomposition. For each level or scale of the decomposition, the energy, entropy and root texture signatures were computed for the approximation and detail wavelet coefficients. The wavelet texture signatures were also computed for the approximation at level 0 (original image). This results in a total of 51 wavelet texture signatures. The analysis was repeated for each of the suggested wavelets.

6.5 Extraction of Multiscale Statistical Texture Features

The third set of texture features are the proposed multiscale statistical texture features. This textural analysis technique is based on a combination of the statistical and multiscale views of texture. Haralick's statistical texture features are used together with a multiresolution analysis to generate a set of multiscale statistical texture features. A formal approach to texture analysis using multiresolution techniques was presented above, in which simple texture signatures were extracted from a multiscale representation of the image. This type of multiscale analysis is now combined with the formal statistical analysis. This provides a complete characterisation of the texture in an image based on the statistical properties of its multiscale representation. The effectiveness of multiscale statistical texture analysis techniques have been recently demonstrated [9].

The wavelet transform is used to map the image samples into a multiscale representation, and the co-occurrence matrices are used to describe the statistical properties of each of the sub-images in the decomposition. A four level wavelet decomposition is performed on all the ROI's in the databases. The co-occurrence matrices are computed for each of the approximation and detail coefficients at each level of the wavelet decomposition and from each of these co-occurrence matrices the eight co-occurrence texture signatures are extracted. This results in a total of 128 co-occurrence texture signatures. There are 96 features extracted from the detail coefficients in the decomposition and 32 from the approximation coefficients. This analysis was also performed for each of the different wavelets discussed.

6.6 Feature Selection and Optimisation

The resulting feature space, incorporating the three groups of texture features, consists of a total of 187 features (8 statistical, 51 wavelet and 128 multiscale statistical). However, although intuitively it would seem that recognition performance is proportional to the dimension of the feature space, this is generally not a valid assumption. As additional features are added to the feature space, recognition performance will at first improve and then eventually deteriorate. This phenomenon is known as the *curse of dimensionality*. The basis for this phenomenon may be explained as follows: given a fixed number of samples any increase in the feature space dimension leads to less points per volume. This leads to the deterioration of the estimates of the class conditional probabilities and results in reduced classification accuracy [50].

Due to the nature of the texture features used, a certain amount of correlation between features and redundancy in the feature space is expected. In the second group of texture features, three different types of wavelet-based texture signatures are evaluated, and there may be correlation between these signatures. Furthermore, many of the features may not contain any significant information, for example, the energy of the approximation coefficients at each level of the decomposition may not contain any significant textural information as most of the information will have been removed due to the iterative low pass filtering in the computation of the wavelet transform. In addition, correlation between the features at different levels of the wavelet decomposition is also possible. Similar redundancy and correlation may occur in the third group of texture features, the multiscale statistical texture features. It is therefore necessary to optimise the feature space, by eliminating redundant and non-discriminatory features. This will shrink the original feature space to a size suitable for computation, and will improve classification accuracy by retaining only the most discriminatory information and deleting irrelevant and redundant information [51].

A feature selection algorithm was employed to select the relevant discriminatory features. A feature selection algorithm consists of two parts, viz. the selection criterion and the search strategy. The selection criterion, or "cost function", is the method of rating the ability of a feature to discriminate between the classes. The search strategy is the method of determining the optimum combination of the rated features.

A probabilistic distance measure was used to rate the efficacy of the features to discriminate between benign and malignant classes of microcalcification. Probabilistic distance measures are based on a measure of the similarity of the class conditional probability density functions $p(x|w_i)$ and $p(x|w_j)$ of the two classes w_i and w_j, integrated over the whole pattern space, using a distance function. These "distances" are strongly correlated with the error of the classifier, with a low classification error corresponding to a large probabilistic distance. The Mahalanobis distance was used as the distance measure [45].

The optimum combination of the features rated using the selection criterion is determined using an algorithmic search strategy. A sequential forward selection (SFS) algorithm was used. This algorithm starts with the best feature, selected based on the rating by the selection criterion, and then tests to see which other feature, excluding the already selected feature scores the highest multivariate criterion. This feature is then included in the reduced feature space. This continues until the required number of features has been selected. In this way, using a probabilistic distance measure as the cost function and a SFS algorithm as a search strategy, a reduced discriminatory feature space could be selected from the original feature space.

6.7 Classification using a K-Nearest Neighbour Classifier

The ability of the extracted texture features to discriminate between benign and malignant classes of microcalcification was demonstrated by using the feature space to classify the microcalcifications using two different classifiers, a K-Nearest Neighbour classifier and an Artificial Neural Network. A nearest neighbour classifier is a non-parametric method to decide to which class a feature sample belongs . Classification of a feature vector is performed by searching for the k closest training vectors based on some distance metric. The test vector is assigned to the class to which the majority of these k nearest neighbours belong [47]. The Euclidean distance measure was used as the distance metric, and in order to prevent the features with the widest variances across the design set from dominating the distance measure, all the features were normalised to have the same variance.

The advantage of the k-nearest neighbour classifier is that it provides an efficient and robust classification scheme without requiring significant initialisation and training time. This allows for the evaluation and comparison of different feature sets.

The classifier was trained and tested using the *leave-out-one* method. This procedure takes the N available samples, trains the network on N-*1* samples and uses the remaining sample as a test case. Classification is continued in this manner until all the N samples have been used as test cases. This training procedure is appropriate for small data sets, as it allows for the maximum size training set, while still allowing for an independent test set. Final performance is reported on the average of the classification results for the N trials.

6.8 Classification using an Artificial Neural Network

A more complex classification scheme, in the form of an artificial neural network, was also implemented. The k-nn classifier is ideal for the comparison of different feature sets, as it is not dependent on training times or initial conditions. However, once an appropriate feature set had been identified, the results obtained using the k-nn classifier were compared with a more complex classifier to see if classification accuracy could be improved.

A feed-forward artificial neural network, with two hidden layers, one input layer and one output layer was designed. The first hidden layer contained 5 neurons, the second 2, and the output layer a single neuron. The input layer merely acts as a buffer to pass on the input data to the first hidden layer. Since simple backpropagation training algorithms suffer from poor training times, a mathematical optimisation technique based on the Levenberg-Marquardt algorithm was used to update the weights and biases during training [55]. Sigmodal transfer functions were used for all neurons in the network, and the network was trained to output a -1 for malignant samples and a +1 for benign samples. The network was trained and tested using the leave-out-one algorithm as described for k-nn classifier.

Chapter 7

Results and Discussion

7.1 Sample Preparation

A set of 97 regions of interest containing clusters of microcalcifications were extracted from the Nijmegen database (29 benign cases and 68 malignant cases). A further set of 83 images was extracted from the LLNL/UCSF database (56 benign cases and 27 malignant cases). Typical ROI's manually extracted from the Nijmegen database are shown in Fig. 7.1 and Fig. 7.2. The original mammogram from which the sample in Fig. 7.2 was extracted is available in Appendix A. Before any features were extracted, the contrast in each of the ROI's was stretched to a grey-level range 0–255. The image texture features were then extracted from each of the sample ROI's. The texture features were also extracted for each of the different wavelets discussed. Thus, the three groups of texture features were extracted from each of the datasets, resulting from each of the experiments performed using the various wavelets.

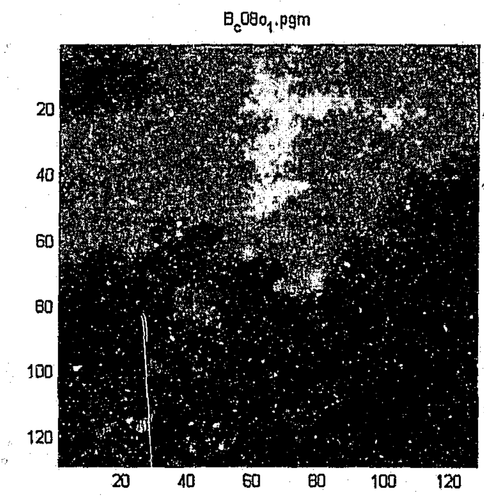

Figure 7.1: An example of a benign ROI extracted from the Nijmegen Database.

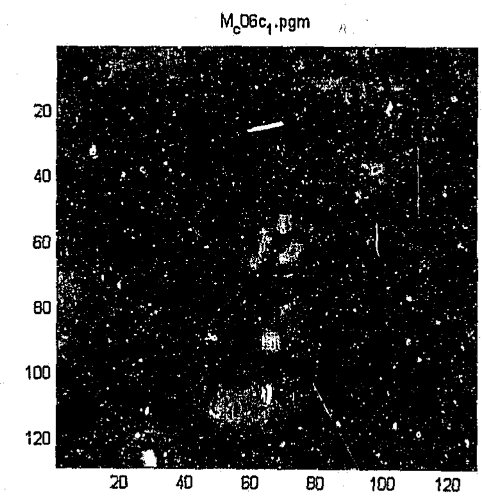

Figure 7.2: An example of a malignant ROI extracted from the Nijmegen Database.

7.2 Feature Selection

The three groups of textural features were extracted from each of the ROI sample images in the two databases, i.e 8 statistical, 51 wavelet and 128 multiscale statistical texture features for each sample image in each database. In the case of the wavelet and multiscale statistical features, different features were also extracted corresponding to each of the four wavelets used, i.e. Biorthogonal and Daubechies 4, 6 and 20. In the case of the statistical and multiscale statistical texture features the displacement vector, d, was also varied between 1 and 9. The feature selection algorithm was then applied separately to each of the groups of texture features, which were extracted from the samples of each dataset, in each of the different configurations of the classification system.

A probabilistic distance measure was used to test the discriminatory ability of the proposed features in each of the groups of texture features. The sequential forward selection algorithm was then applied to the rated features to select the optimum combination of a subset of between 8 and 12 features. In this way redundancy and correlation was removed from each of the groups of features, and only a subset of discriminatory features was retained.

For the statistical texture features, three of the features, viz. *cluster shade*, *cluster prominence* and the *information measure of correlation* were selected as prominent features in all cases, i.e. for all displacement vectors used, and for both databases. This implies that in all cases these features are rated by the feature selection algorithm as the most important when differentiating between the texture in benign and malignant ROI's. There was also found to be no difference between the different displacement vectors, i.e. the same three features were selected as prominent regardless of the displacement vector used in the calculation of the co-occurrence matrix.

For the group of wavelet-based texture signatures, the *energy* texture feature proved to be dominant, across all the wavelets used and both databases. This was expected as Laine and Fan [36] compared energy and entropy features and found the energy features to be more suitable for texture classification. The selected energy features were from all levels of the decomposition and a range of detail coefficients.

For the multiscale statistical texture features, multiscale features based on *cluster shade*, *cluster prominence* and the *information measure of correlation* tended to dominate the optimised feature sets. This was expected, as these features were found to be significant at a single scale. *Contrast* also proved to be a discriminatory feature.

The feature selection algorithm was also applied to the combined feature set, i.e. the full feature space comprising the statistical features, the wavelet-based texture signatures and the multiscale statistical texture features. For this combined feature set, similar results to those obtained for the individual feature sets were obtained. Thus, the wavelet energy features, and the multiscale statistical *cluster shade* and *prominence*, *contrast* and *correlation* measures were most frequently selected.

An interesting observation is that the same features proved to be prominent in both databases used. This is encouraging as it highlights the robustness of each of the groups of texture features.

The detailed results of the feature selection algorithm are available in Appendix C.

These reduced feature sets, for each of the four groups of texture features, i.e. the statistical texture features, the wavelet-based texture signatures, the multiscale statistical texture features and a combined set of features, are used as input to the classifiers.

7.3 Classification using a K-Nearest Neighbour Classifier

The k-nn classifier is ideal for the comparison of different types of feature sets, as it requires no significant training or initialisation time. In contrast, in a more complex classification scheme, like a neural network, there is some randomness associated with the initialisation and training of the classifier. A k-nn classifier therefore provides a

Displacement	Classification Accuracy
1	75.26% (3)
2	75.26% (8)
4	74.29% (3)

Table 7.1: Classification results for statistical texture features for various displacement vectors (Nijmegen Database).

robust and efficient method for the comparison of different feature sets.

Tables 7.1, 7.2, 7.3 and 7.4 present a summary of the results. The classification accuracies for each of the three groups of texture features are shown. For each group of texture features, between 8 and 12 optimum features, identified by the feature selection algorithm, were used for classification. Tables 7.1 and 7.2 depict the results obtained for the Nijmegen digital mammogram database. Table 7.1 shows the results obtained for the statistical texture features for various displacement vectors, while Table 7.2 shows the results obtained for the wavelet and multiscale statistical texture features, using different wavelets. A combined feature set is also included, which contains the combination of the wavelet, multiscale statistical and the set of statistical features corresponding to a displacement vector of 1. Tables 7.3 and 7.4 depict the corresponding results for the LLNL/UCSF database. The number of features used in each case is also shown in brackets. Typically best classification accuracy was achieved when using between 5 and 10 of the selected features. As additional features are added, classification accuracy will at first improve and then either level off or decrease. This is shown in Fig. 7.3.

From Tables 7.1 and 7.3 it can be seen that there is no significant difference between the results obtained when using different displacement vectors for the computation of the co-occurrence matrices. This was also observed during feature selection, i.e. the same features were selected regardless of the displacement vector used. Thus, statistical features calculated using a displacement vector of 1 were used for the combined feature set.

From Tables 7.2 and 7.4 it can be observed that in all cases the multiscale statistical features perform as well or better than the wavelet-based texture signatures. In

Wavelet Type		Feature Set	
	Wavelet	Multiscale Statistical	Combined Feature Set
Biorthogonal 2.8	72.17% (8)	82.47% (10)	82.47% (10)
Daubechies 4	76.29% (7)	76.29% (5)	76.29% (5)
Daubechies 6	74.23 (5)	77.32% (8)	77.32% (8)
Daubechies 20	74.23 (6)	79.39% (10)	79.39% (10)

Table 7.2: Classification results for Nijmegen Database

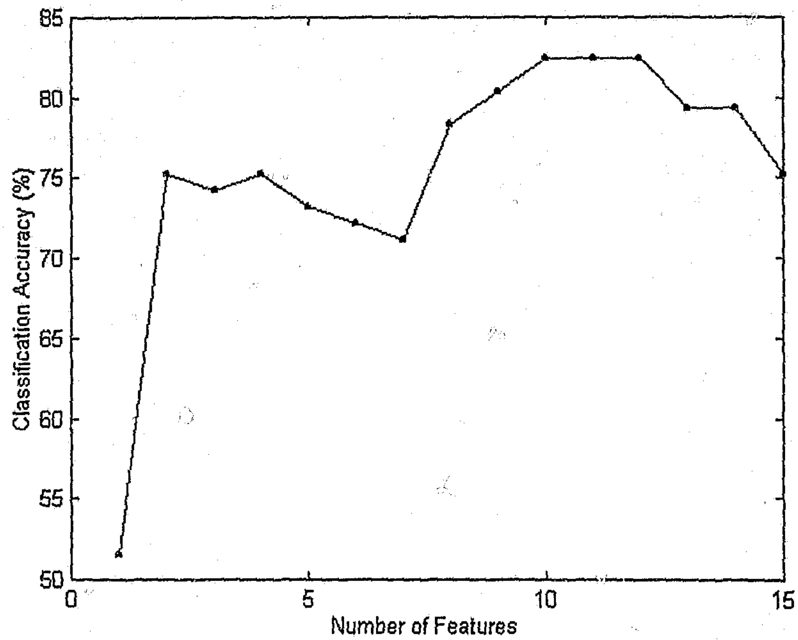

Figure 7.3: Classification performance versus feature set dimensionality.

Displacement	Classification Accuracy
1	74.7% (3)
2	75.9% (3)
6	72.3% (7)

Table 7.3: Classification results for statistical texture features for various displacement vectors (UCSF/LLNL Database)

Wavelet Type		Feature Set	
	Wavelet	Multiscale Statistical	Combined Feature Set
Biorthogonal 2.8	73.49% (3)	79.52% (3)	80.72% (9)
Daubechies 4	72.29% (9)	77.11% (9)	79.52% (11)
Daubechies 6	71.08% (4)	78.31% (12)	79.52% (11)
Daubechies 20	71.08% (4)	83.13% (6)	83.13% (6)

Table 7.4: Classification results for LLNL/UCSF Database

addition there is no real significant difference between the results obtained using the different wavelets. Furthermore, the results obtained for the combined feature sets are only marginally better than the results obtained for the multiscale statistical texture features. This was expected as the multiscale statistical texture features encompass the multiscale properties of the wavelet-based texture features as well as the statistical properties of the original co-occurrence texture features. Therefore, not much improvement in the classification accuracy is obtained when using the combined feature set.

7.4 Classification using an Artificial Neural Network

Improved classification accuracy was achieved using an artificial neural network. The combined feature set for each of the wavelets was used as input to the neural network. Ten of the twelve features extracted from the combined feature set using the feature selection algorithm were used as discriminatory features for classification. The classification accuracy of the neural network is shown in Tables 7.5 and 7.6, for each of the wavelets and for each database.

From Tables 7.5 and 7.6 and Tables 7.2 and 7.4 it can be seen that the neural network outperforms the k-nn classifier for all the wavelets used. Thus, the use of a more

Wavelet Type	Classification Accuracy
Biorthogonal 2.8	93.8% (10)
Daubechies 4	85.5% (10)
Daubechies 6	92.7% (10)
Daubechies 20	94.8% (10)

Table 7.5: Classification results for the neural network for each of the wavelets used (Nijmegen Database).

Wavelet Type	Classification Accuracy
Biorthogonal 2.8	89.1% (10)
Daubechies 4	91.5% (10)
Daubechies 6	90.3% (10)
Daubechies 20	93.9% (10)

Table 7.6: Classification results for the neural network for each of the wavelets used (LLNL/UCSF Database).

complex classifier provides an improvement in the classification accuracy. It can also be seen that here too there is no significant difference between the different types of wavelets used.

Further insight into the discriminatory ability of the texture features can be seen in the confusion matrices shown in Tables 7.7 and 7.8.

The neural network was trained to produce a -1 output for benign cases and a +1 output for malignant cases. The threshold was set at -0.5, i.e. outputs above -0.5 are classified as malignant and outputs below and including -0.5 are classified as benign. This threshold was arrived at during experimentation, and was chosen to minimise the false negative rate, i.e. the ratio of the number of malignant cases misclassified to the total number of malignant cases. The confusion matrices in Tables 7.7 and 7.8 reflect the true positive and false positive rates of the system. The true positive rate is defined as the ratio of the number of malignant cases correctly classified to the total number of malignant cases. The false positive rate is defined as the ratio of the number of benign cases incorrectly classified to the total number of benign cases. On the confusion matrices the columns reflect the predicted classes and the rows the expected classes (i.e. the biopsy proven ground truth). Thus, the second row in each matrix depicts the true positive rate, which lies between 85% and 98%, with an average of 95%. The false positive rate can be observed in the first row of each

Biorthogonal 2.8		
	Benign	*Malignant*
Benign	25	4
Malignant	2	66
Daubechies 4		
	Benign	*Malignant*
Benign	17	12
Malignant	2	66
Daubechies 6		
	Benign	*Malignant*
Benign	23	6
Malignant	1	67
Daubechies 20		
	Benign	*Malignant*
Benign	25	4
Malignant	1	67

Table 7.7: Confusion matrices for Nijmegen classification results.

Biorthogonal 2.8		
	Benign	*Malignant*
Benign	49	7
Malignant	2	25
Daubechies 4		
	Benign	*Malignant*
Benign	53	3
Malignant	4	23
Daubechies 6		
	Benign	*Malignant*
Benign	49	7
Malignant	1	26
Daubechies 20		
	Benign	*Malignant*
Benign	52	4
Malignant	1	26

Table 7.8: Confusion matrices for LLNL/UCSF classification results.

matrix. A high true positive rate is achieved, at the expense of some false positives. However, an important observation is that there is a low false negative rate.

Therefore, for both databases and all the wavelets used, the neural network outperforms the k-nn classifier. From the confusion matrices it can also be observed that a high true positive rate is achieved, and the combination of the three groups of texture features provide significant discriminatory ability for the discrimination between benign and malignant classes of microcalcifications.

7.5 Summary of the Results

The most important results obtained may be summarised as follows:

7.5.1 Feature Selection

Statistical Texture Features

The system was implemented in a variety of different configurations and from each of these experiments the most discriminatory features were selected using the feature selection algorithm. From the results of the feature selection algorithm the following was established:

- The statistical texture features are roughly independent of the displacement vector used in the computation of the co-occurrence matrix.
- Three features emerged as prominent in all the configurations of the system, viz. *cluster shade*, *prominence* and *correlation*.
- The results were roughly independent of the dataset used.

Multiscale Texture Features

- As expected the *energy* signatures emerged as prominent.

- The selection results are roughly independent of the wavelet used.
- The selection results are roughly independent of the dataset used.

Multiscale Statistical Techniques

- The multiscale features corresponding to the fixed-scale statistical features emerged as prominent.
- The selection results are roughly independent of the dataset used.

Combined Feature Set

- The features selected corresponded to those selected in the individual feature sets.

7.5.2 Classification using a K-nn Classifier

- The classification results are roughly independent of the dataset used.
- The results are roughly independent of the wavelet used.
- The multiscale statistical texture features outperform the other texture features.
- Not much improvement in the classification accuracy is achieved when using a combined feature set.

7.5.3 Classification using an Artificial Neural Network

- The neural network outperforms the k-nn classifier in all configurations.

Chapter 8

Conclusions and Scope for Future Work

8.1 Conclusions

In this investigation various texture analysis techniques have been used to differentiate between benign and malignant classes of microcalcification clusters contained in digitised mammograms. The hypothesis is that there is some form of textural information surrounding the tissue containing the clusters of microcalcifications, and that this information can be used as an aid in the diagnosis of the microcalcification.

Microcalcifications are an indication of an active degenerative process, and various classification systems have been developed that are capable of analysing the properties of segmented microcalcifications and using this information as a basis for discriminating between benign and malignant processes. The conjecture here is that there is some form of underlying textural information contained in the tissue surrounding the microcalcifications. The biological process that causes the microcalcifications is contained in the tissue and the microcalcifications are merely the result of this process. An attempt is made to analyse the actual tissue, i.e. the entire region of interest (ROI) surrounding the microcalcification, and use this as a basis for discriminating between benign and malignant processes.

Our analytical techniques are based on a collection of image texture analysis methods. Three different types of image texture features were compared in the analysis, to determine their suitability for the classification of a ROI containing microcalcifications. These include: statistical texture features based on the co-occurrence matrix, wavelet-based multiscale texture signatures and our proposed multiscale statistical texture signatures, which are based on a multiscale and statistical analysis of image texture.

Based on this investigation it would seem that texture methods are an appropriate tool for the analysis of the regions of interest containing microcalcifications, as the texture features successfully distinguish between the two classes of microcalcifications. Furthermore, the group of multiscale statistical features outperform the other features, and are capable of successfully discriminating between benign and malignant regions of interest containing microcalcifications.

The co-occurrence matrix is used to extract statistical texture information, and the wavelet transform is used to decompose the sample ROI images into a multiscale representation. These two techniques form the basis of the first two groups of texture features, and are combined to form the third group of texture features.

During the feature selection phase it was found that for each of the categories of texture features similar features were selected regardless of the wavelet used, or image database used. This seems to indicate that there are certain features in each category that contain more discriminatory ability. Each of the subsets of optimised features were then used as input to a k-nn classifier. In all cases, the multiscale statistical features performed best. In addition a combined feature set incorporating all of the features only marginally improved the classification accuracy, which indicates that the multiscale statistical features encompassed the other types of texture features. The combined group of texture features was also tested using an artificial neural network, which gave better classification results than the k-nn classifier.

Thus, in conclusion, based on these experimental results, it appears that there is some form of textural information surrounding microcalcification areas in digitised mammograms. Furthermore, it is possible to characterise this textural information using texture analysis techniques. Therefore, texture features can be extracted from these areas and used to successfully discriminate between the benign and malignant processes that exist in these areas. In addition from the range of texture features evaluated, the proposed multiscale statistical texture features seem to be the most valuable.

However, although the robustness of the texture features was verified by testing the system on two datasets, both datasets contain a limited number of samples. Further investigation will still need to be conducted using a larger mammographic database to increase the statistical significance of the results.

It should also be noted that the aim of this investigation was to evaluate the ability of texture analysis techniques to classify microcalcifications. The effectiveness of these techniques was successfully demonstrated using a number of different types of texture analysis techniques. However, it would also have been possible to test the efficacy of the features using a different classifier and/or feature selection technique. In fact, any number of different classification configurations could have been implemented which may, or may not have improved the classification accuracy of the system. However, this does not effect the significance of the results, which demonstrate the diagnostic value of texture methods for the classification of microcalcifications.

8.2 Scope for Future Work

There is definite scope for the further development of this classification system. Firstly, the system could be augmented with additional features useful for distinguishing between classes of microcalcifications. The proposed range of texture techniques provide valuable discriminatory information. However, it might be useful to incorporate a number of features based on the properties of the actual microcalcifications, e.g. their size, orientation shape roughness etc. Patient demographics might also provide important information useful for diagnosis and could also be included.

Secondly, this system is a single component of a wider digital mammography workstation currently under development at the Department of Electrical Engineering, University of the Witwatersrand. It is envisaged that this classification system will be incorporated into a full mammographic workstation capable of providing extensive diagnostic information to the radiologist.

The complete mammographic workstation will be configured to work as follows. Firstly, the system will be provided with digitisation means, for acquiring mammographic samples. It will also be provided with a microcalcification detection scheme, for automatically detecting microcalcifications. This detection scheme will be coupled to the classification scheme proposed in this study. Furthermore, a system for detecting and classifying mass lesions will also run in tandem with the microcalcification system. The workstation thus provides the radiologist with a suite of tools for the detection and diagnosis of abnormal lesions. Therefore, there is scope for the development of a means of interfacing this system with the complete mammographic workstation.

Finally, in order to thoroughly test the significance of the classification system, it would have to be set up as an active on-line system in parallel to existing diagnostic procedures. The system would then provide an indication of the likelihood of benignancy or malignancy of an area containing microcalcifications. These results could then be verified or rejected based on actual clinical biopsy results. This information would then be used in the continued on-line training and verification of the proposed system.

Bibliography

[1] M. Moskowitz, "Mammography to screen asymptomatic women for breast cancer", *American Journal of Radiology*, vol. 143, pp. 457-459, 1984.

[2] R. E. Bird, T. W. Wallace and B. C. Yankaskas, "Analysis of cancers missed at screening mammography", *Radiology*, vol. 184, pp. 613-617, 1992.

[3] M. L. Giger, "Current issues in CAD for mammography", *Proc. of the 3rd International Workshop on Digital Mammography*, Elsevier Science, 1996.

[4] E. L. Thurnfjell, K. A. Lernevall and A. A. S. Taube, "Benefit of double reading in a population-based mammography screening program", *Radiology*, vol. 191, pp. 241-244, 1994.

[5] F. M. Hall, J. M. Storellas, D. Z. Silverstone and G. Wyshak, "Nonpalpable breast lesions: recommendations for biopsy based on suspicion of carcinoma at mammography", *Radiology*, vol. 167, pp. 353-358, 1988.

[6] R. M. Haralick, K. Shanmugam and I. Dinstein, "Texture features for image classification", *IEEE Trans. on Systems Man Cybern.*, vol. 3, no. 6, pp. 610-621, 1973.

[7] A. P. Dhawan, Y. Chitre, C Kaiser-Bonasso and M. Moskowitz, "Analysis of mammographic microcalcifications using grey-level image structure features", *IEEE Tran. Med. Imag.*, vol. 15, no. 3, pp. 246-259, June 1996.

[8] C. M. Kocur, S. K. Rogers, L. R. Myers, T. Burns, M. Kabrisky, J. W. Hoffmeister, K. W. Bauer and J. M. Steppe, "Using neural networks to select wavelet features for breast cancer diagnosis", *IEEE Eng. Med. Biol.*, pp. 95-102, May/June 1996.

[9] G. Van de Wouwer, Wavelets for multiscale texture analysis, PhD thesis, Dept. Physics, University of Antwerp, 1998.

[10] M. L. Giger, Computer-Aided Diagnosis, RSNA Categorical Course in Physics: Technical Aspects of Breast Imaging, Radiological Society of North America, Ed. A. G. Haus and M. J. Yaffe, pp. 283-298, 1993.

[11] N. Karssemeijer, M. Thijssen, J. Hendricks and L. van Erning (eds.), Digital Mammography Nijmegen 1998, Kluwer Academic Publishers, Dordrecht, 1998.

[12] K. Doi, M. L. Giger, R. M. Nishikawa and R. A. Schmidt (eds.), Digital Mammography '96, Elseveir Science, 1996.

[13] L. V. Ackerman, A. N. Mucciardi, E. E. Gose and F. S. Alcorn, "Classification of benign and malignant tumours on the basis of 36 radiographic properties", *Cancer* vol. 31, pp. 342-352, February 1973.

[14] J. A. Baker, P. J. Kornguth, J. Y. Lo, M. E. Williford and C. E. Floyd, "Breast cancer: prediction with artificial neural network based on BI-RADS standardised lexicon", *Radiology*, vol. 196, pp. 817- 812, 1995.

[15] J. A. Baker, P. J. Kornguth, J. Y. Lo and C. E. Floyd, "Artificial neural network: Improving the quality of breast biopsy recommendations", *Radiology*, vol. 198, pp. 131-135, 1996.

[16] J. Kilday, F. Palmieri and M. D. Fox, "Classifying mammographic lesions using computerised image analysis", *IEEE Trans. Med. Imag.* vol. 12, no. 4, pp. 664 - 669, 1993.

[17] S. Pohlman, K. A. Powell, N. A. Obuchowski, W. A. Chilcote and S. Grundfest-Broniatowski, "Quantitative classification of breast tumours in digitised mammograms", *Med. Phys.* 23 (8), pp. 1337-1345, 1996.

[18] Z. Hou, M. L. Giger, C. J. Vyborny, U. Bick and P. Lu, "Analysis of spiculation in the computerised classification of mammographic images", Med. Phy. 22(10), pp. 1569-1579, 1995.

[19] M. L. Giger, C. J. Vyborny and R. A. Schmidt, "Computerised characterisation of mammographic masses: analysis of spiculation", *Cancer Letters*, vol. 77, pp. 201-211, 1994.

[20] L. Shen, R. M. Rangayyan and J. E. L. Desautels, "Application of shape analysis to mammographic calcifications", *IEEE Trans. Med. Imag.*, vol. 13, no. 2, pp. 263-274, June 1994.

[21] Y. Wu, M. L. Giger, K. Doi, C. J. Vyborny, R. A. Schmidt, C. E. Metz, "Artificial neural networks in mammography: application to decision making in the diagnosis of breast cancer", *Radiology*, vol. 187, pp. 81-87, 1993.

[22] R. M. Nishikawa, M. L. Giger, K. Doi, C. J. Vyborny, R. A. Schmidt, C. E. Metz, Y. Wu, F. Jiang, Z. Hou, P. Lu, W. Zhang, T. Ema, U. Bick, J. Papaioannou, R. H. Nagel, "Computer-aided diagnosis of masses and clustered microcalcifications from digitised mammograms", *SPIE* vol. 1905, pp. 422-432, 1993.

[23] Y. Jiang, R. M. Nishikawa, D. E. Wolverton, C. E. Metz, M. L. Giger, R. A. Schmidt, C. J. Vyborny, K. Doi, "Malignant and benign clustered microcalcifications: automated feature analysis and classification", Radiology, vol. 189, pp. 671-678, 1996.

[24] F. Aghdasi, R. K. Ward and B. Palcic, "Classification of mammographic microcalcifications", *CCECE/CCGEI '93*, pp. 1196-1199, 1993.

[25] E. A. Patrick, M. Moskowitz, V. T. Mansukhani and E. I. Gruenstein, "Expert learning system network for the diagnosis of breast calcifications", *Investigative Radiology*, vol. 26, pp. 534-539, 1990.

[26] Z. Hou, M. L. Giger, C. J. Vyborny, D. E. Wolverton, R. A. Schmidt and K. Doi, "Computer-aided diagnosis: classification of mammographic mass lesions," *Digital Mammography '96*, K. Doi, M. L. Giger, R. M. Nishikawa and R. A. Schmidt (eds.), Elseveir Science, pp. 207-210, 1996.

[27] D. Brzakovic, X. M. Luo and P Brzakovic, "An approach to automated detection of tumours in mammograms", *IEEE Trans. Med. Imag.*, vol. 9, no. 3, pp. 233-241, 1990.

[28] C. J. Burdett, H. G. Longbotham, M. Desai, W. B. Richardson and J. F. Stoll, "Nonlinear indicators of malignancy", *SPIE* vol. 1905, pp. 853-860, 1993.

[29] A. Grossman and J. Morlet, "Decompositions of hardy functions into integrable wavelets of constant shape", *SIAM J. Math. Anal.* vol. 15, pp. 723-736, 1984.

[30] I. Daubechies, "Orthonormal bases of compactly supported wavelets", *Comm. Pure. Appl. Math.*, vol. 41, pp. 909-996, 1988.

[31] S. Mallat, "A Theory for multiresolution signal decomposition: the wavelet representation", *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 11, no. 7, pp. 647-693, 1989.

[32] I. Daubechies, Ten Lectures on Wavelets, Capital City Press, Montpellier, Vermont, 1992.

[33] G. Strang and T. Nguyen, Wavelets and Filter Banks, Wellesley-Cambridge Press, 1996.

[34] T. H. Koornwinder (ed.), Wavelets: An elementary treatment of the theory and applications, World Scientific, 1993.

[35] M. Misiti, Y. Misiti, G. Oppenheim and J. Poggi, Wavelet Toolbox User's Guide, The Mathworks, 1996.

[36] A Laine and J Fan, "Texture classification by wavelet packet signatures", *IEEE Trans. Patt. Anal. Machine Intell.*, vol. 15, no. 11, pp. 1186 - 1191, November 1993.

[37] IEEE Standard 610.4-1990, IEEE Standard Glossary of Image Processing and Pattern Recognition Terminology, IEEE Press, New York, 1990.

[38] A. K. Jain, Fundamentals of Digital Image Processing, Prentice Hall, 1998.

[39] M. Tuceryan and A. K. Jain, Handbook of Pattern Recognition and Computer Vision, Chapter 2.1: Texture Analysis, World Scientific, 1993.

[40] R. Jain, R. Kasturi and B. G. Schunck, Machine Vision, McGraw-Hill, 1995.

[41] P. Scheunders, S. Livens, G. Van de Wouwer, P. Vautrot and D. Van Dyck, "Wavelet based texture analysis", to be published as an invited paper in *International Journal Computer Science and Information management*, December 1997.

[42] T. Chang and C.C. Jay Kuo, "Texture analysis and classification with tree-structured wavelet transform", *IEEE Trans. Image Proc.*, vol. 2, no. 4, pp. 429 - 441, October 1993.

[43] Y. Chitre, M-band wavelet discrimination of natural textures with application to the classification of mammographic microcalcification, PhD Thesis, Division of Research and Advanced Studies of the University of Cincinnati, 1996.

[44] R. N. Strickland and H. Hahn, "Wavelet transforms for detecting microcalcifications in mammograms", *IEEE Trans. Med. Imag.*, vol. 15, no. 2, pp. 218-229, 1996.

[45] P. Devijver and J. Kittler, Pattern Recognition: A Statistical Approach, Prentice-Hall, Englewood Cliffs N.J., London, 1982.

[46] V. Vapnick, Estimation of Dependences Based on Empirical Data, Springer-Verlag, New York, 1982.

[47] R. O. Duda and P. E. Hart, Pattern Classification and Scene Analysis, J. Wiley and sons, 1973.

[48] R. Schalkhoff, Pattern Recognition: statistical, structural and neural approaches, John Wiley and Sons, New York, 1992.

[49] J. T. Tou and R. C. Gonzalez, Pattern Recognition Principles, Addison-Weseley, Reading, MA, 1974.

[50] K. Fukunaga, Introduction to Statistical Pattern Recognition, Academic Press, London, 2nd Edition, 1990.

[51] K. R. Castleman, Digital Image Processing, Prentice-Hall, Englewood Cliffs N.J., 1979.

[52] R. Hecht-Nielson, Neurocomputing, Addison Wesley, Reading, 1990.

[53] M. Caudill, Neural Networks Primer, Miller Freeman Publications, San Francisco, 1989.

[54] A. Jain and D. Zongker, "Feature selection - evaluation, application and small sample performance", *IEEE. trans. Pat. Anal. Mach. Intell.*, vol. 19, no. 2, pp. 153-158, 1997.

[55] M. T. Hagan and M. Menhaj, "Training feedforward networks with the Marquardt Algorithm," *IEEE Trans. Neural Networks*, vol. 5, no. 6, pp. 989-993, 1994.

[56] Department of Radiology, University Hospital Nijmegen, PO Box 9101, The Netherlands.

[57] Lawrence Livermore National Laboratories (LLNL) / University of California at San Francisco (UCSF) Radiology Department, e-mail: mammo-db-help@llnl.gov.

Appendix A

Example Mammographic Image

An example of a typical mammogram is shown in Fig. A.1. The example mammogram is a cranio-caudal image of the breast obtained from the Nijmegen Digital Mammogram Database [56]. A malignant cluster of microcalcifications is highlighted on the image. This cluster of microcalcifications was extracted and used by the classification system. The extracted cluster is shown in Fig. 7.2.

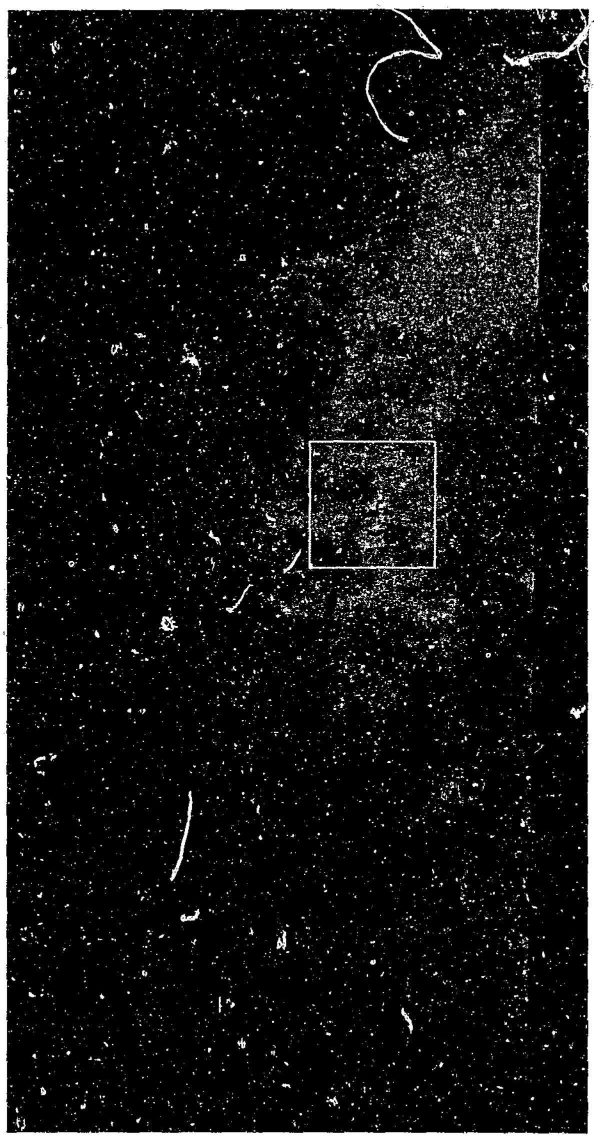

Figure A.1: Example mammographic image, including highlighted microcalcification area.

Appendix B

Statistical Texture Features

Haralick [6] defines a set of co-occurrence texture features which may be used to extract the statistical nature of the texture available in the image, via the co-occurrence matrix. Many of these measures relate to the specific textural characteristics of the image, such as homogeneity, contrast, entropy, energy and the presence of organised structure in the image. Others relate to the complexity and nature of the grey-tone transitions which occur in the co-occurrence matrix. Further information regarding the interpretation of these measures is available in [6]. In the definitions, $H(y_i, y_j, d)$ refers to the co-occurrence matrix calculated from each of the sample ROI's, where: y_i and y_j are the grey level pairs separated by a displacement vector d. These statistical texture features are defined as follows:

$$Contrast = \sum_{i,j=0}^{n} (y_i - y_j)^2 H(y_i, y_j, d) \tag{B.1}$$

$$Angular\ Second\ Moment = \sum_{i,j=0}^{n} [H(y_i, y_j, d)]^2 \tag{B.2}$$

$$Entropy = -\sum_{i,j=0}^{n} H(y_i, y_j, d) \log [H(y_i, y_j, d)] \tag{B.3}$$

$$Inverse\ Difference\ Moment = \sum_{i,j=0}^{n} \frac{1}{1+(y_i-y_j)^2} H(y_i, y_j, d) \tag{B.4}$$

$$Maximum\ Probability = max_{i,j}\left[H(y_i, y_j, d)\right] \tag{B.5}$$

$$Cluster\ Shade = \sum_{i,j=0}^{n} (i - M_x + j - M_y)^3 H(y_i, y_j, d) \tag{B.6}$$

where:

$$M_x = \sum_{i,j=0}^{n} y_i H(y_i, y_j, d)$$

$$M_y = \sum_{i,j=0}^{n} y_j H(y_i, y_j, d)$$

$$Cluster\ Prominence = \sum_{i,j=0}^{n} (i - M_x + j - M_y)^4 H(y_i, y_j, d) \tag{B.7}$$

$$Information\ Measure\ of\ Correlation = \frac{(Entropy - H_{xy})}{max\left[H_x, H_y\right]} \tag{B.8}$$

where:

$$H_{x=} - \sum_{i=0}^{n} S_x(y_i) \log\left[S_x(y_i)\right] \ ,$$

$$H_{y=} - \sum_{j=0}^{n} S_y(y_j) \log\left[S_y(y_j)\right] \ ,$$

$$H_{xy} = -\sum_{i,j=0}^{n} H(y_i, y_j, d) \log\left[S_x(y_i) S_y(y_j)\right]$$

and

$$S_x(y_i) = \sum_{j=0}^{n} H(y_i, y_j, d)\ ,$$

$$S_y(y_j) = \sum_{i=0}^{n} H(y_i, y_j, d)$$

Appendix C

Feature Selection Results

The feature selection algorithm was applied to each of the groups of features, i.e. to
The feature selection algorithm was applied to each of the groups of features, i.e. to
the statistical, multiscale, multiscale statistical texture features and the combined feature set, incorporating all three groups of features. The classification system was also implemented in a number of variations, i.e. for a range of different wavelets, and different displacement vectors. The system was also tested using two different datasets. The feature selection algorithm was therefore applied to each of the different feature sets extracted from each database in each of these configurations. A summary of these results is presented graphically in Fig. C.1. These two plots show the full combined feature set on their *x-axes*, and the number of times a particular feature was selected on the *y-axes*.

The full 187 features are shown on the *x-axes*. These are arranged as follows: The first 51 features correspond to the multiscale features, the remaining 136 features are the statistical and multiscale statistical texture features. The individual feature sets are then ordered as follows. The first 17 features are the *root* wavelet signatures, which include the signature extracted from the original image as well as the signatures extracted from each level of 4-level wavelet decomposition, for each of the approximation and three detail subimages. The next 17 features are the *energy* signatures, ordered similarly, followed by 17 *entropy* signatures. Thereafter,

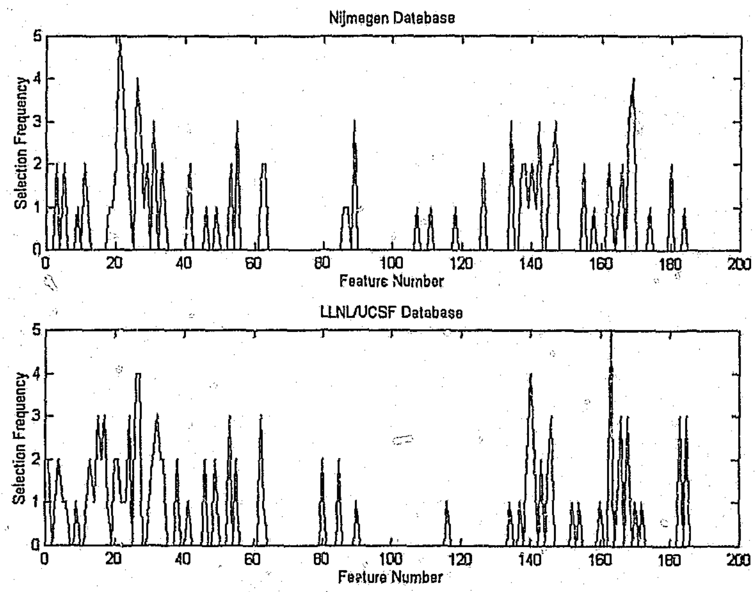

Figure C.1: Feature selection results for the combined feature set, for each of the two databases.

the statistical, and multiscale statistical texture features are arranged in a similar way. For each of the 8 statistical texture features, the feature extracted from the original image, corresponding to the fixed-scale statistical analysis is positioned first, followed by the statistical features extracted form the multiscale representation of the image, as described above. The order, in groups of 17, of the combined feature set is therefore as follows:

1-17	Root signature
18-34	Energy signature
35-51	Entropy signature
52-68	Contrast
69-35	Angular Second Moment
86-102	Entropy
103-119	Inverse Difference Moment
120-136	Max. Probability

137-153 Cluster Shade

154-170 Cluster Prominence

171-187 Information Measure of Correlation

From these figures it can be seen that for both datasets similar features emerge as prominent, which seems to indicate that the features are independent of the dataset used. In addition, from the activity around certain areas of the plots, the prominent features can be seen. Specifically, the prominence of the multiscale energy signatures, which correspond to feature numbers 18 - 34 can be seen, similarly, the prominence of the multiscale statistical features, namely: contrast (52 - 68), cluster shade (137 - 153), cluster prominence (154 - 170) and correlation (171 - 187), are also visible on the plot.