

Johnson & Wales University ScholarsArchive@JWU

Engineering Studies Faculty Publications and
Creative Works

College of Engineering & Design


2000

Application of Wavelets and Principal Component Analysis in Image Query and Mammography

Sol Neeman Ph.D.

Johnson & Wales University - Providence, sneeman@jwu.edu

Follow this and additional works at: https://scholarsarchive.jwu.edu/engineering_fac

 Part of the [Computer Engineering Commons](#), [Electrical and Computer Engineering Commons](#), [Engineering Science and Materials Commons](#), [Mechanical Engineering Commons](#), [Operations Research, Systems Engineering and Industrial Engineering Commons](#), and the [Other Engineering Commons](#)

Repository Citation

Neeman, Sol Ph.D., "Application of Wavelets and Principal Component Analysis in Image Query and Mammography" (2000). *Engineering Studies Faculty Publications and Creative Works*. 1.
https://scholarsarchive.jwu.edu/engineering_fac/1

This Dissertation is brought to you for free and open access by the College of Engineering & Design at ScholarsArchive@JWU. It has been accepted for inclusion in Engineering Studies Faculty Publications and Creative Works by an authorized administrator of ScholarsArchive@JWU. For more information, please contact jcastel@jwu.edu.

**APPLICATION OF WAVELETS AND PRINCIPAL COMPONENT ANALYSIS IN IMAGE
QUERY AND MAMMOGRAPHY**

BY

SOL NEEMAN

A Dissertation submitted in partial fulfillment

of the requirements for the degree of
DOCTOR OF PHILOSOPHY

IN

APPLIED MATHEMATICAL SCIENCES,

Specialization in: Computer Science

University of Rhode Island

2000

Copyright by Sol Neeman

Abstract

Breast cancer is currently one of the major causes of death for women in the U.S. Mammography is currently the most effective method for detection of breast cancer and early detection has proven to be an efficient tool to reduce the number of deaths. Mammography is the most demanding of all clinical imaging applications as it requires high contrast, high signal to noise ratio and resolution with minimal x-radiation. According to studies [36], 10% to 30% of women having breast cancer and undergoing mammography, have negative mammograms, i.e. are misdiagnosed. Furthermore, only 20%-40% of the women who undergo biopsy, have cancer. Biopsies are expensive, invasive and traumatic to the patient. The high rate of false positives is partly because of the difficulties in the diagnosis process and partly due to the fear of missing a cancer. These facts motivate research aimed to enhance the mammogram images (e.g. by enhancement of features such as clustered calcification regions which were found to be associated with breast cancer) , to provide CAD (Computer Aided Diagnostics) tools that can alert the radiologist to potentially malignant regions in the mammograms and to develop tools for automated classification of mammograms into benign and malignant classes. In this paper we apply wavelet and Principal Component analysis, including the approximate Karhunen Loeve transform to mammographic images, to derive feature vectors used for classification of mammographic images from an early stage of malignancy.

Another area where wavelet analysis was found useful, is the area of image query. Image query of large data bases must provide a fast and efficient search of the query image. Lately, a group of researchers developed an algorithm based on wavelet analysis that was found to provide fast and efficient search in large data bases. Their method overcomes some of the difficulties associated with previous approaches, but the search algorithm is sensitive to displacement and rotation of the query image due to the fact that wavelet analysis is not invariant under displacement and rotation. In this study we propose the integration of the Hotelling transform to improve on this sensitivity and provide some experimental results in the context of the standard alphabetic characters.

Acknowledgment

First, I would like to thank my advisor, Professor Bala Ravikumar from the University of Rhode Island, Computer Science department, for the support and encouragement he provided in my work on this dissertation. I particularly benefitted from his suggestion to study wavelet theory which has led to the work on combining the Hotelling transform in Image Query. The many discussions we had together were very fruitful and provided valuable input for my research.

Second, I would like to thank Professor Nathan Intrator from Brown University who guided me in the research on wavelet analysis and suggested the problem of classification of mammographic images. Professor Intrator introduced me to the areas of high dimensional signal analysis and advanced methods in wavelet packet analysis. He provided endless support and encouragement during our joint work. I profited immensely from my association with him. He helped in clarifying many ideas in the area of multidimensional signal representation, feature extraction and classification and his insights into these ideas has had a deep influence on my thinking. I was very fortunate in working with him.

Last but not least, I would like to give my special thanks to my wife, Amy, who provided all the support possible in completing this thesis, in proofreading and advising on the style of this thesis. Without her I could not complete this thesis.

Dedicated to My Beloved Daughters, Danielle and Dahlia

Table Of Contents

A	LIST OF FIGURES	ix
1.	INTRODUCTION	1
1.1	Wavelets: A Brief Historical Review	1
1.2	Wavelet Analysis	4
1.3	Wavelets and Compression	7
1.4	Wavelets and Signal Denoising	8
1.5	Wavelet and Multiresolution	8
1.6	Wavelets and Wavelet Packets	9
1.6.1	Best Basis and Joint Best Basis	10
1.7	Principal Component Analysis	11
1.8	Image Query, a Brief Introduction	12
1.9	Wavelet Analysis and Mammography	14
1.9.1	Review of Selected Previous Work	15
1.10	Overview of the Thesis	17
2.	MATHEMATICAL PRELIMINARIES	20
2.1	A Brief Review of Wavelet Theory	21
2.1.1	Linear Space and Wavelets	21
2.1.1.1	Creating an orthonormal basis by defining a wavelet subspace	22

2.1.2	The Dilation Equation and the Construction of the Scaling and Wavelet functions	23
2.1.3	The Fast Wavelet Transform	24
2.1.4	One-level and multilevel decomposition of a signal	25
2.1.5	The Reconstruction Process	25
2.1.6	Wavelets and Filter Banks	26
2.1.7	Vanishing Moments of a Wavelet Function	30
2.1.8	1-D Wavelet Analysis	30
2.1.9	2-D Wavelet Analysis	32
2.1.10	2-D wavelet analysis of images	34
2.1.11	De-noising of 2-dimensional images	34
2.2	Wavelet Packets	35
2.2.1	Orthonormal Bases and Information Cost Functions	37
2.2.2	The Best Basis Algorithm	38
2.2.3	The Joint Best Basis (JBB)	42
2.3	Principal Component Analysis (Karhunen Loeve Transform)	44
2.3.1	Application of the KLT in Geometric Manipulation of Images	46
2.3.2	Application of the KLT in aligning an image along its principal axis	47
2.4	The Approximate Karhunen Loeve Transform	48
2.4.1	Biplot of data using the first two coordinates in a feature vector	50
2.5	Feature Extraction and Classification	51

2.5.1	Measures of Energy Distributions	51
2.5.1.1	Transform Coding Gain(TCG)	52
2.5.1.2	Accumulation of Variance	53
2.5.2	Compression and Approximation Error	53
2.5.3	Fisher's Linear Discriminant Analysis (LDA)	54
3.	METHODOLOGY	57
3.1	Combining the Hotelling Transform in Image Query	57
3.1.1	Wavelet Based Approach in Image Query	57
3.1.2	Aligning an image along its principal axis	58
3.2	Enhancement and Classification of Mammographic Images	58
3.2.1	Mammographic Data Base	58
3.2.2	General framework in representing and analyzing an image	59
3.2.3	Indicators for Breast Cancer in Mammographic Images	60
3.2.4	Image Enhancement as a Preprocessing Step for classification	61
3.2.5	Feature Extraction	63
3.2.6	Justification for Using Feature Vectors Based on Wavelet Packet Coefficients	65
3.2.7	Application of the Joint Best Basis to Mammographic Images	66
3.2.8	Comparison of different bases using their accumulated variance	67
3.2.9	Classification Framework for Mammographic Images	69
3.2.10	Feature Vectors	69

3.2.10.1	Discriminating Power of Coefficients in a Feature Vector	72
3.2.11	Classifiers	72
3.2.11.1	K-Nearest Neighborhood (k-nn) Classifier	73
3.2.11.2	Classification of Mammograms Using the K-nn Algorithm	73
3.2.11.3	Fisher's Linear Discriminant Analysis(LDA)	74
3.2.12	Classification Using the <i>Joint Best Basis</i>	75
3.2.12.1	Classification using the <i>common Joint best basis</i>	76
3.2.12.2	Classification using <i>two Joint best bases</i>	77
3.2.12.3	Classification using individual feature vectors	77
3.2.13	Application of the Approximate Karhunen Loeve Transform to Mammographic Images	77
3.2.13.1	Plotting data with respect to the first two coordinates of the approximate KL basis	78
4.	EXPERIMENTAL RESULTS	80
4.1	Effects of image enhancement	80
4.2	Comparing the Accumulated Variance in Various Bases	82
4.3	Classification results	86
4.3.1	Classification Results Using Fisher's Linear Discriminant	86
4.3.1.1	Feature Vectors Based on the Common Joint Best Basis of Enhanced Mammograms	87
4.3.1.2	Feature Vectors Based on the Common Joint best basis of Unprocessed Images	88
4.3.1.3	Distribution of Wavelet Packet Coefficients in the Common Joint Best Basis	88

4.3.2	Classification Results Using a Set of Two Joint Best Bases	90
4.4	Feature Vectors Based on Variance Values	91
4.4.1	Classification of Mammograms Using the K-nn Algorithm	94
4.4.2	Feature Vectors Based on the Transform Coding Gain(TCG)	96
4.4.3	Feature vectors based on the accumulated variance	98
4.4.3.1	Evaluation of the results	100
4.4.4	Classification Results Using the Approximate KLT	102
4.4.5	Biplot using the approximate KLT basis and the Joint best basis	104
4.4.5.1	Comparing variance values and accumulated variance values as feature vectors	104
4.5	Summary	105
5.	COMBINING THE HOTELLING TRANSFORM IN IMAGE QUERY	107
6.	CONCLUSION	120
6.1	Summary of the Results	120
6.2	Classifiers as a second opinion for radiologist	121
6.3	Future Research	121
B	REFERENCES	123
C	BIBLIOGRAPHY	125

List Of Figures

1.1	50Hz sine wave 'buried' in noise. Top panel: the signal in the time domain, bottom panel, the magnitude of its FFT.	2
1.2	A 50Hz signal with an abrupt discontinuity. Top panel, the signal in the time domain. Bottom panel shows the magnitude of its FFT. The location of the abrupt change cannot be deduced from the FFT information	3
1.3	Morlet wavelet and its translated versions	5
1.4	Morlet wavelet and its scaled versions	6
2.1	db6 4 qmf filters. Left panel from top: Scaling filter, LP decomposition filter, LP reconstruction filter, transfer modulus of LP filter. Right column from top: HP decomposition filter, HP reconstruction filter, transfer modulus of HP filter.	29
2.2	A signal with second derivative discontinuity. Top panel- signal, mid panel-second detail coeff's, bottom panel-first level detail coeff's.	31
2.3	Wavelet analysis of a signal composed of white noise. From top to bottom panel: analyzed signal, detail coeff's at level 4, 3,2,1 respectively.	32
2.4	De-noising of a signal. Top panel- noisy electrical signal, bottom- de-noised version.	33
2.5	Two levels of nonstandard decomposition of the image 'woman'	35
2.6	De-noising of images. Left top-original signal, right top-original signal with added noise, bottom left-de-noised signal.	36
2.7	First step in best basis search: Mark all bottom nodes in the binary tree representing the entropy values of each crystal(represented by a node)	41
2.8	Second step in search for a best basis: compare the entropy of parent and its children and mark all nodes (marked with an asterisk) with lower cost, starting from the bottom.	42
2.9	Final step in search for a best basis: select topmost marked nodes (enclosed by rectangulars) to get the best basis.	43

2.10	Application of the KL to a 2-dimensional image results in a rotated version of the image along its principal axis.	47
2.11	Illustration of Fisher Linear Discriminant as a classifier. The projection of the samples onto Fisher's LD are used for classification.	55
3.1	The Karhune Loeve coordinates for a population of 2-dimensional vectors consisting of two classes	65
3.2	Classification results using the Knn classifier for 10 experiments, each with 10,20,30 coefficients: top panel-average error, mid panel-sensitivity, bottom - specificity.	74
4.1	Unprocessed and enhanced images and their frequency spectrum	81
4.2	histogram of pixel densities for enhanced and unprocessed images	82
4.3	accumulation of variance in the Joint best basis for enhanced and unprocessed image. x axis: coefficient number.(64 coefficients for segments of 8x8 pixels)	83
4.4	Accumulation of variance in different bases for the first enhanced benign mammogram. x axis: coefficient number(64 coefficients for segments of 8x8 pixels)	84
4.5	Accumulation of variance in different bases for the first unprocessed benign mammogram. x axis: coefficient number (64 coefficients for segments of 8x8 pixels)	85
4.6	Accumulation of variance in different bases for the first enhanced malignant mammogram. x axis: coefficient number (64 coefficients for segments of size 8x8 pixels)	86
4.7	Accumulation of variance in different bases for the first unprocessed malignant mammogra(x axis: coefficient number. 64 coefficients for segments of size 8x8)	87
4.8	Section (magnified) of the accumulated variance in the PC basis and the Joint bb for the first benign and malignant images using segment size of 16x16 pixels.	88
4.9	Histogram plot of Fisher Linear Discriminant values for enhanced mammograms	89
4.10	Histogram plot of Fisher's Linear Discriminant for the unprocessed images	90

4.11	Distribution of the first and second wavelet packet coefficients with largest discriminating power for enhanced segments using the common Joint best basis	91
4.12	Distribution of the first two wavelet packet coefficients with the largest discriminating power for unprocessed images.	92
4.13	Scatter plot of training segments using two Joint best bases	93
4.14	Scatter plot of test segments using two Joint best bases	94
4.15	Mean Square Error of ensemble of benign and malignant segments, reconstructed using 20% of the coefficient.	95
4.16	Classification results of 10 experiments using the K-nn classifier for 10,20 and 30 coefficients. Top panel-average error, mid panel-sensitivity, bottom panel-specificity.	96
4.17	Discriminating power plot using the Transform Coding Gain as feature vectors.	98
4.18	Discriminating power plot of the accumulated variance in the PC basis and the Joint best basis for the first option.	99
4.19	Discriminating power plot of the accumulated variance in the PC basis and the Joint best basis for the second option	100
4.20	Scatter plot of mammograms using option 1 with 6 coefficients of the accumulated variance	101
4.21	Scatter plot of mammograms using option 2 with 42 coefficients of the accumulated variance.	102
4.22	Classification results of 50 experiments using 10 coefficients and the Knn classifier	103
4.23	Classification results of 50 tests using the Approximate KL Transform as feature vector	104
4.24	Biplot of mammograms using the first coefficients in the approximate KL basis and the Joint best basis	105
5.1	Sensitivity of wavelet coefficients to displacement and rotation	119

Chapter 1

Introduction

In this chapter, we provide a brief historical review of wavelets and their applications in signal processing including a few illustrative examples. We begin by pointing out the drawbacks of the Fourier transform and the windowed Fourier transform in the analysis of certain signals, factors that motivated the development of wavelet analysis. We continue with a brief and qualitative description of what wavelet analysis is, its use in compression and de-noising and its multiresolution properties which makes it attractive for certain applications in computer science. We briefly describe some of the important tools used in this study e.g. the Best Basis, the Joint Best Basis and the Karhunen Loeve transform. All these subjects will be presented formally and in detail in chapter 2. We also provide a brief introduction to image query and a review of previous work on wavelet analysis in mammography. We end the chapter with an overview of the thesis.

1.1 Wavelets: A Brief Historical Review

In this section we provide a brief historical review of wavelets based mainly on [29]. The area of wavelets is relatively new. Its roots go back at least a century to the work of Karl Weierstrass, in which he describes a family of functions constructed by superposition of scaled copies of a base function. These functions were fractiles, being everywhere continuous but nowhere differentiable. In many disciplines there is a need to analyze signals or data which usually come in a time series form, e.g., acoustic data, speech or music, seismic information, various medical signals, images, and financial data. These signals and data have to be cleaned up from noise, encoded, compressed or analyzed for the detection of specific patterns. The classical tool for signal and time series data analysis is Fourier analysis. The Fourier transform breaks the signal into its frequency components and provides spectral information. In many cases, the Fourier transform provides the necessary information one looks for. If a certain dominant

frequency is hidden in noise for example, Fourier analysis can, under appropriate conditions, detect that frequency in a straightforward way. Fig 1.1 shows a 50 Hz sine wave 'buried' in noise. It is hard if not impossible to detect the existence of this sinusoid in the time domain. But as can be seen, it is clearly visible in the frequency domain. The problem with Fourier analysis is that the frequency coefficients are an 'average' over the whole time interval. Sines and cosines are local in the frequency domain but global in the time domain. In other words, in the transformation to the frequency(amplitude) domain, the time information is lost. The magnitude spectrum of a signal does not provide any information relating the frequency component to time. If, for example, there is an abrupt change in the signal, Fourier analysis is not suited to detect (with reference to the time domain) such an abrupt change. Fig 1.2 shows a signal with an abrupt change. The location of the abrupt change cannot be deduced from its frequency spectrum. In chapter 2 we provide some illustrative examples of the capabilities of wavelet analysis in detecting discontinuities, even minute ones.

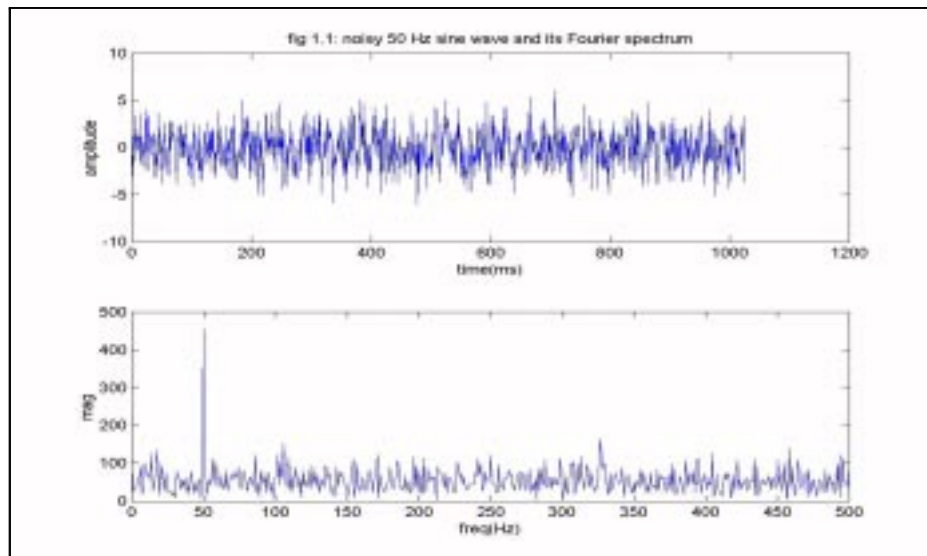


Figure 1.1. 50Hz sine wave 'buried' in noise. Top panel: the signal in the time domain, bottom panel, the magnitude of its FFT.

In 1946, Dennis Gabor adapted Fourier analysis to perform local analysis. He developed a technique called 'windowing', in which a small section of the signal is analyzed at a time. This is the *Short-Time*

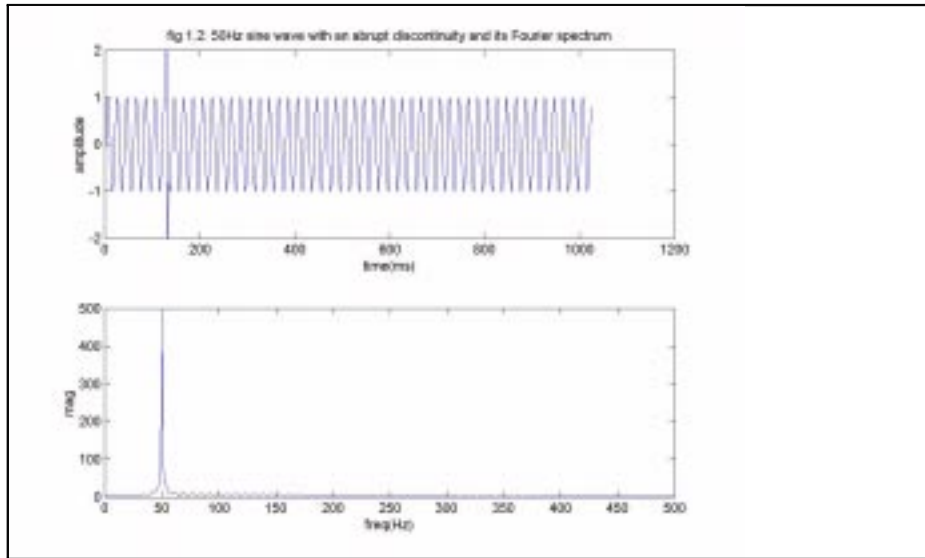


Figure 1.2. A 50Hz signal with an abrupt discontinuity. Top panel, the signal in the time domain. Bottom panel shows the magnitude of its FFT. The location of the abrupt change cannot be deduced from the FFT information

Fourier Transform(STFT). It maps a signal into a two dimensional function of time and frequency. This analysis may provide a solution in the analysis of certain signals. In a chirp signal, for example, the frequency of the signal increases continuously. The instantaneous frequency, however, is well defined. The Fourier spectrum of a chirp signal is very broad and does not reveal the well defined instantaneous frequency. The *STFT* does show that at each point of time there is a dominant frequency component[14]. The Short-Time Fourier Transform has some drawbacks as its analysis is somewhat arbitrary. The window size it uses may not be the optimal one for the analysis of the signal. Different ways of 'chopping' up a signal may yield different results, and therefore it has to be tailored to the particular signal under analysis [14]. In addition, the STFT does not provide information on the very low frequency components of the signal due to the fact that the window size is limited. In general, functions obey a version of the uncertainty principle. Sharp localization in time and frequency are mutually exclusive.. To gain information on low frequencies we need larger time intervals.

Wavelets were the next step in improving the tool for time frequency analysis of a signal. Wavelets use a flexible window (scale) and thus can provide information on the low frequency components of the

signal (which require large time intervals) and also provide localized information about high frequencies (requiring short time intervals).

Back in 1909, Alfred Haar constructed an orthonormal system of functions with compact support which could provide local information in the analysis of a signal. The system he constructed is called the Haar basis and it provides the simplest wavelet among the family of wavelets developed so far.

The origin of the term *wavelet* comes from the field of seismology, where it describes the disturbance generated by a sharp seismic impulse that propagates upward. Later work by Morlet et al showed how these seismic wavelets can be described with the Gabor functions. Morlet showed that any arbitrary function can be analyzed with a set of analyzing functions composed of a single scaled and translated mother wavelet. The notion of a single basis function used to build an entire basis using its scaled and translated versions is in the heart of wavelet theory. It was further developed into a sound theory called *multiresolution analysis* by Yves Meyer and Stephen Mallat. Figures 1.3 and 1.4 illustrate the scaled and translated versions of the Morlet wavelet. It is the local support of the basis function that enables local analysis of the signal. Shifted versions of the wavelet function enable the analysis of all the sections in a signal and scaled versions of the wavelet function provide a wide range of frequency information about the signal.

Other researchers followed by developing richer families of wavelet functions with various properties such as vanishing moments, support and regularity (Coifman, Daubechies and many others). For example, the wavelet function developed by Daubechies provided a set of smooth basis functions which are orthonormal and have compact support. In chapter 2, we illustrate the derivation of the set of basis functions associated with Daubechies [4].

1.2 Wavelet Analysis

Wavelets are waveforms with limited duration (compact support) unlike the sine and cosine waveforms which are the basis functions in Fourier analysis. They are usually asymmetric and irregular in

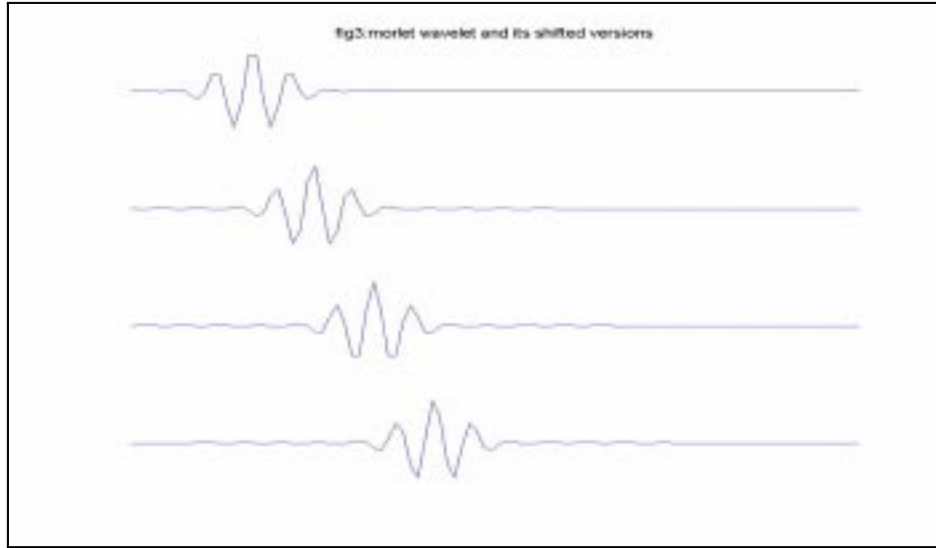


Figure 1.3. Morlet wavelet and its translated versions

shape (unlike sine and cosine waveforms). The rapid changes in the waveform of a wavelet function accounts for its ability to analyze those portions of a signal which contain sharp changes. Wavelet analysis has a continuous form, the *Continuous Wavelet Transform (CWT)*, in which the change in scale and shift of the wavelet waveform are continuous, and a discrete form, the *Discrete Wavelet Transform (DWT)*, in which the change in scale is performed on dyadic intervals (the ratio of intervals is a power of 2). There is a one-dimensional wavelet analysis and a two-dimensional wavelet analysis. In the DWT, the signal passes through a pair of filters (decomposition filters) that represent the scaling and wavelet filters, one a low pass and the other a high pass filter. The output of both filters is then down sampled. The result (which has the same length as the original signal) is composed of an 'average' and 'details' of the signal (due to the application of a pair of low pass and high pass filters). In wavelet analysis, the process of applying the pair of filters is repeated on the average portion of the result. In *wavelet packet* analysis the process is applied to both the average and the detail coefficients, resulting in a finer resolution of the frequency space. If the decomposition and reconstruction filters satisfy certain conditions (which will be described in chapter 2), then the reconstruction of the signal is perfect.

One step of wavelet decomposition is illustrated below:

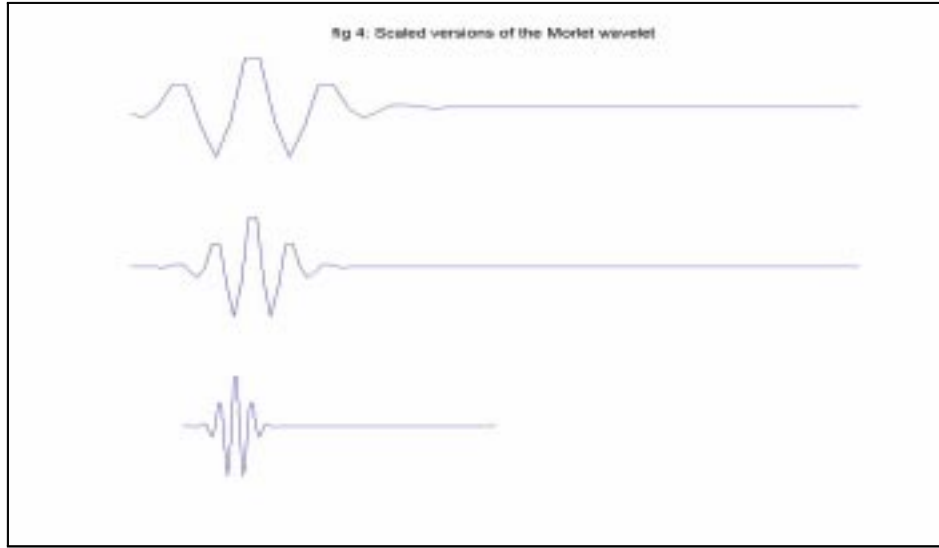


Figure 1.4. Morlet wavelet and its scaled versions

$$\begin{array}{l}
 \nearrow \text{HighPass Filter} \longrightarrow (\downarrow \text{downsampling}) \longrightarrow (N/2) \text{ detail} \\
 \text{Signal} \\
 (N \text{ samples}) \\
 \searrow \text{LowPass Filter} \longrightarrow (\downarrow \text{downsampling}) \longrightarrow (N/2) \text{ approx.}
 \end{array}$$

The reconstruction of the signal from its wavelet coefficients consists of similar steps but in reverse order. First the average coefficients and the detail coefficients are upsampled; then they are passed through a pair of reconstruction filters and added together:

$$\begin{array}{l}
 (N/2) \text{ detail coeff's} \rightarrow (\uparrow \text{upsampling}) \rightarrow \text{HighPass Filter} \searrow \\
 \text{Reconstructed} \\
 \text{Signal} \\
 (N/2) \text{ approx. coeff's} \rightarrow (\uparrow \text{upsampling}) \rightarrow \text{HighPass Filter} \nearrow
 \end{array}$$

In two-dimensional wavelet analysis, which is applied to analyze images, the wavelet function is composed of a scaling function and three wavelet functions. The result of one step analysis is an average of the image (a smaller, lower resolution replica) and three sets of details: horizontal, vertical and diagonal details. Two-dimensional analysis will be presented in more detail in chapter 2.

1.3 Wavelets and Compression

Images of interest contain certain regularity and are not composed of pure noise. This means that there is a certain amount of correlation in the pixels' intensities. In general, if the pixels are highly correlated, then there is a lower rank description of the image which captures most of the variance of the coordinates. In transform coding, we seek a basis in which the coordinates are either uncorrelated or less correlated so that most of the variation takes place in many fewer coordinates. If we can find such a transformation, we then can represent the image in fewer coordinates. In such a case we achieve data compression. The efficacy of a transformation depends on how much energy compaction is provided by the transform. The Transform Coding Gain (TCG) [26], is one way of measuring the amount of energy compaction achieved by an orthonormal transformation. The wavelet transform is found to have a high value of TCG. While other transforms like the Discrete Cosine Transform may achieve a better compression ratio, they do not have the adaptive property of the wavelet transform. In other transforms, the rule of zeroing small coefficients is applied evenly and globally over all detail coefficients while the wavelet transform is adaptive in this respect. It allows preserving small coefficients which account for 'important' minute features that we do not want to be lost or distorted in the transformation.

Recently, the FBI had chosen the wavelet/scalar quantization(WSQ) algorithm over JPEG to store its data base of fingerprints[30]. The FBI has a collection of 30 million sets of fingerprints that have to be stored in an easily accessible form. Each fingerprint requires about 0.6 megabytes of memory. Although JPEG can provide a compression ratio of 15:1 to 20:1, the blocking artifacts distort the features which are important for identification. These features are the 'minute' ridge endings and bifurcations (which form permanently in childhood). The ability of wavelet analysis to provide a high level of compression while preserving these minute features made it a winner [30]. As the amount of data today becomes enormous, the need for compression is more crucial. Networks have to handle enormous amounts of data in the form of text, sound, images or video clips. Satellite imagery and medical images require enormous amounts of memory and compression is crucial for saving memory and for efficient and fast image retrieval. A single

24 bit color picture with 256x256 pixels would require more than 0.2 megabytes of memory. Compression increases the throughput of the network and, for satellite transmission, compression would greatly reduce cost and decrease transmission time.

1.4 Wavelets and Signal Denoising

One of the applications of wavelets in the area of signal processing is noise removal. When decomposing a signal, most of the wavelet coefficients are small or zero. This is the property that enables wavelets to compress signals. In addition, since the noise is usually white noise, its energy is spread over the whole spectrum. In the wavelet coefficients the noise energy will be concentrated in the detail coefficients. Various 'thresholding' techniques were developed to remove most of the noise from a signal while decreasing the loss in the signal[21,33]. In chapter 2, we provide some illustrative examples of the de-noising power of wavelet analysis.

1.5 Wavelet and Multiresolution

Wavelet analysis provides a structure that represents a signal or image at different levels of approximations[29]. The mathematical framework of wavelet bases is that of a nested linear spaces $V^0 \subset V^1 \subset V^2 \subset \dots$. The basis functions of these linear spaces are called scaling functions. In addition to the scaling function we define the *wavelet spaces*, W^j , which satisfies:

$$V^{j+1} = V^j \oplus W^j$$

where \oplus denotes the orthogonal sum operator. The wavelet functions at level j , W^j , are the orthogonal complement of V^{j+1} in the space V^j . The basis functions in *Wavelets* are the basis functions that constitute the wavelet space. This mathematical framework enables representing an image in a multiresolution mode, i.e. the image is represented with different levels of approximation. This representation

can be useful, for example, in image editing where it is important for the user to be able to do changes at a coarser level in certain areas of the image and make finer changes in other areas.

Multiresolution representation of images has found many applications. These include: image query, interactive painting, texture mapping (portion of a texture can be defined in a higher resolution), data bases that provide a set of images coalesced into a single multiresolution image, virtual reality, and games among other applications. The data structure used for multiresolution imaging is the quadtree which enables the implementation of an efficient algorithm for reconstruction of various parts of the image at a desired resolution.

1.6 Wavelets and Wavelet Packets

The wavelet transform has a continuous form and a discrete form just like the Fourier transform and the Discrete Fourier Transform. We will limit the discussion to the Discrete Wavelet Transform (DWT). It was discovered that the DWT can be implemented using dyadic scales and positions (powers of 2) without compromising accuracy. This provides an efficient algorithm called the *fast wavelet transform*. The idea is to apply a set of two analyzing filters, high pass and low pass filters, each followed by a down sampling operation. The output of the low pass filter provides information on the low frequency components of the signal while the output of the high pass filter contains information on the high frequency components of the signal. In chapter 2, we provide the implication of these convolution-downsampling operations in the frequency domain to see how the spectrum of the original signal relates to that of the decimated outputs of the filter banks.

In wavelet multilevel analysis, the decomposition process is iterated over the approximation coefficients, resulting in the *wavelet decomposition tree*. Signals and images of interest have some structure. Normally most of the energy would be concentrated in the low frequency range. As a result, the multilevel wavelet analysis provides 'layers' of information on the signal. The first level of detail coefficients will provide information on the highest frequency components; the second level of detail coefficients will

provide information on the next band of frequencies and so on. At the last level we find the approximation coefficients.

In wavelet packet analysis, the decomposition process is applied to both the approximation and the detail coefficients. The frequency spectrum of the signal is analyzed into a finer partition and that provides a richer family of basis functions. This leads us to the *best basis* algorithm developed by Coifman and Wickerhauser, which selects a basis from among a family of bases which is optimal with respect to some information cost function. The best basis algorithm will be described in details in chapter 2. This is important in the search algorithm for the 'best basis' as will be described in chapter 2.

1.6.1 Best Basis and Joint Best Basis

As was mentioned, wavelet packet analysis provides a rich family of basis functions (using a certain wavelet function). This family offers a large number of bases, some of them orthogonal bases. The family of basis functions can be searched for a 'best basis' with regard to some information cost function. We will use entropy as the cost function, as it will provide a good measure of the compression ability of the basis. The fast algorithm developed by Coifman and Wickerhauser searches the best basis with computational complexity of $ON(\log N)$ steps where N is the length of the data. This best basis is the basis that would have the smallest reconstruction error among all possible bases in the family of basis functions when the signal is reconstructed from a subset of its wavelet packet coefficients (which normally are taken to be a subset with the largest magnitude).

The concept of best basis was extended to a family of signals or vectors by Wickerhauser. Given an ensemble of vectors, a wavelet packet analysis can be applied to the ensemble. The wavelet packet coefficients can be used to derive what is called the *Joint best basis*. This basis best describes the ensemble (with respect to some information cost function) among all possible bases in the joint packet table. It is important to note that a signal can be analyzed using various wavelets and thus a best basis or a joint best basis can be derived for each wavelet. Details on the joint best basis will be provided in chapter 2.

1.7 Principal Component Analysis

Principal component analysis (PCA), also known as the Karhunen-Loeve transform(KLT), Factor Analysis and Hotelling transform, was first introduced by Pearson in 1901 and developed by Hotelling in 1933. It is the best known tool for multivariate analysis, and as computers gained high computational power, it became more and more useful and popular, especially in statistical analysis[13]

Given a set of observations (or population of vectors) each with a dimension d , the various parameters (coordinates) maybe correlated and so the data is redundant. This implies that the data can be represented in lower dimensional coordinates. PCA provides a transformation in which the original set of observations is transformed to a new set of decorrelated coordinates. This results in a new distribution of the variance of the population where most of the variance is concentrated in a fewer $d' \ll d$ coordinates (called the principal components). This reduction of dimensionality can be useful in terms of computation and in terms of being able to describe some properties or features with a smaller number of parameters. The Transform Coding Gain (TCG) or the accumulated variance is a measure that can be used to compare the distribution of the variance in the new coordinates to that in the old coordinates . If the population satisfies a multivariate normal distribution, PCA will provide the highest TCG among all orthogonal transformations [26].

The KLT is also useful in compression due to the fact that most of the variance will be accounted by fewer coordinates. It can also be used as a tool to detect clusters in a population [13]. If the population can be represented in a small number of dimensions, then the PCA will find it so that for a high dimensional data, we can get a two dimensional representation using the first two largest Principal Components. Whether this would enable detecting clusters in the population depends on the features distinguishing the classes.

With respect to classification, if the statistics of the training population and the test population are similar, we can use the KLT to derive a transformation matrix from the training data and apply it to the test data for classification.

As the transformation matrix in PCA consists of the eigenvectors of the covariance matrix of the population, the transformation is data dependent (unlike the DCT which uses fixed basis functions). For that reason, in applications of data compression, if the statistics of a channel are nonstationary, the transformation matrix has to be recomputed and has to be sent to the receiver. This high overhead of computation may limit the KLT in those applications [26].

The Karhunen-Loeve transform has also a geometric application which will be used in this thesis [9]. In this application, the coordinates of the pixel's image (rather than the intensity of the pixels) are used to form the vector population (for 2-D images the dimension of the vector population would be 2). When the KLT is applied over this vector population, the geometric effect is the rotation of an image along a direction in which it seems most elongated (called the 'principal axis' of the image). This application is useful in image recognition and will be used in this thesis to improve the sensitivity to rotation of a wavelet based approach in image query, as will be described in the next section.

1.8 Image Query, a Brief Introduction

Many fields such as graphic design, architecture, TV production, multimedia, art history, and satellite imaging offer large data bases with thousands of images the user can access. These data bases have to offer a fast, efficient way of searching for an item out of thousands of images. Traditional methods for searching these data bases fail due to their large size, as it is very hard to locate a query image among several thousands of target images[29].

One common strategy used in searching a query image in a large data base includes indexing the images in the data base with keywords, but this approach has many difficulties from the user's perspective which make it unacceptable for large data bases[29].

In another strategy called the 'content-based image query', the query is provided by the user either as the output of some electronic device such as a low-resolution output of a scanner or video camera or a rough sketch painted by the user. For example, a graphic designer may want to search for a specific image

within its own data base or an external data base containing high resolution images by providing a low resolution version of the image. We now mention some of the difficulties associated with this strategy. The query image may be very different from the target image; so the comparison should tolerate such difference. Also, if the query image is the output of some electronic device, it may suffer artifacts due to color shift, poor resolution, etc., and if the image is hand painted, it is subject to perceptual errors. As a result of these difficulties, the use of L^1 and L^2 metrics in the query process are not effective. These metrics cannot accommodate the large differences and distortion between the query image and the target images. Recently a group of researchers [12] developed a multiresolution image query algorithm based on wavelet analysis, which greatly improves the success rate and speed of the query search compared to other content based image query systems. The metric used in this approach is based on truncated, quantized wavelet coefficients of the images. A subset of the wavelet coefficients is then chosen (those with large magnitude) and their values quantized to ± 1 . Then, in the comparison process, only the significant wavelet coefficients are used and, due to the quantization, the metric accommodates differences in their values. A data structure also was developed for the fast computation of this metric. As an example, the query process of a 128x128 image on a database of 20,000 images takes under 1/2 a second compared to 14 minutes when the L^1 metric is used.

Some of the previous approaches to content based image query, such as the ones using a color histogram, are not sensitive to image displacement and rotation. IBM developed a system which is commercially available and allows the query process to be based on the color, composition, texture, shape feature and dominant edges. Another approach called query by visual example, performs edge extraction on the sketched input and uses these edges to search the data base.

Compared to these approaches, the multiresolution approach provides some advantages, such as the ability to compare a query image and a target image regardless of the resolution with which each is represented. Still it suffers from high sensitivity to displacement and rotation of query image relative to the target image as wavelet coefficients are not invariant under displacement and rotation of an image. We will combine the Karhunen Loeve transform in the query process to improve this sensitivity. A detailed

description of this approach including some experimental results within the limited context of alphabetic characters and including the improvement on the sensitivity to rotation and displacement can be found in a published paper found in chapter 5.

1.9 Wavelet Analysis and Mammography

Mammography is currently the most effective method for detection of breast cancer, and asymptomatic screening with mammography can reduce mortality from breast cancer by 20-40% [36]. It is also the most demanding of all clinical imaging applications as it requires high contrast, high signal to noise ratio and resolution with minimal x-radiation. The success rate in the diagnosis depends on the skills of the radiologist and the visual inspection of the mammogram.

Breast cancer is one of the major causes of death for women in the U.S. Early detection using mammography has proven to be an efficient tool to reduce the number of deaths. Still 10% to 30% of women who have breast cancer and undergo mammography have negative mammograms [1]. Furthermore, only 20%-40% of the women who undergo biopsy have cancer. Biopsies are expensive, invasive and traumatic to the patient. This motivated the development of CAD (Computer Aided Diagnosis) tools as well as research on automated classification of mammograms to help the radiologist decrease the percentage of false negatives and increase the positive predictive value in the diagnosis process.

The high rate of false positives is partly due to the difficulties in the diagnosis process and partly due to the fear of missing a cancer. In a general screening population, less than 1% of the cases will have cancer. The rate of success in the diagnosis of mammograms depends greatly on image quality, signal to noise ratio and the resolution of the image. Research in developing digital imaging is in progress [5], but currently mammogram images are based on the conventional screen film. Screen film mammograms are limited in their dynamic range, and also by low contrast resolution, film noise and film processing artifacts [5] which are in part responsible for the high rates of false negatives and false positives. In about two thirds of the false negative cases, the radiologist failed to detect cancer that was evident in

retrospective possibly due to image quality, eye fatigue and distraction by other image features. These facts motivate research aimed at enhancing the mammogram images (e.g. by enhancement of features such as clustered calcification regions which were found to be associated with breast cancer), providing CAD (Computer Aided Diagnostics) tools that can alert the radiologist to potentially malignant regions in the mammograms, and developing tools for automated classification of mammograms into benign and malignant classes. Compression of mammographic images that greatly reduce the number of coefficients needed to represent the image, while still keeping the features that discriminate between malignant and benign mammograms (whether visible or not visible to the radiologist), are therefore important.

In this thesis we apply a few variants of the *joint best basis* algorithm and the *approximate KLT*, which combines the joint best basis derived from the wavelet packet of an ensemble of segments and principal component analysis to mammographic images. We use these tools for image enhancement and classification of mammograms to benign and malignant classes.

1.9.1 Review of Selected Previous Work

Most of the previous work on mammographic images consists of image enhancement (to improve the detection of calcification points and masses) and automated detection of calcification clusters and masses. Microcalcifications are an important indicator of malignancy as 30-50% of the malignant mammograms contain calcification points [34]. Fewer studies have been done in automated classification of mammographic images.

A variety of approaches have been used to enhance mammographic features, to detect and localize suspicious areas, and to classify mammographic images including expert-based systems (that use texture analysis), neural networks, Fourier analysis and wavelet analysis. The Department of Radiology, Kurt Rossmann Laboratories conducted intensive research in various breast imaging projects. In [38] a computer program has been developed which can locate clustered microcalcifications on mammograms. In this project, the mammogram is first enhanced using linear filtering to improve the signal to noise ratio of

microcalcifications regions. Then using gray-level thresholding techniques, potential areas of microcalcifications are isolated and indicated by the program. The rate of true-positive cluster detection reported using 60 mammograms reached 87%. In [39] a computerized scheme for the detection of masses is being developed. The method employed in this project is based on the deviation from the normal architectural symmetry of the right and left breasts to detect potential candidate masses. In an evaluation study, using 154 pairs of clinical mammograms, a sensitivity of 95% was achieved with an average of 2.5 false-positive detections per image. (Note: not all masses are malignant). In [41] an automated computerized classification of clustered microcalcification was developed. The method uses features of individual microcalcification (thickness, volume, area, and shape) as well as features of the cluster itself (number of calcification points in the cluster, area and shape of the cluster). These projects were integrated in [40] into an intelligent mammography workstation, and 8,035 cases were analyzed in the first two years of implementation. It was found that *out of the total 34 cancers found in the population, 23 were detected by the workstation* (16 of 23 cases contained masses and 7 of 23 cases contained clustered microcalcifications). In [42], the authors use a morphological based algorithm to detect microcalcification points. The output of the system developed is a mammogram with circled microcalcification. The interpretation and diagnosis are left for the mammographer. In [35], several wavelets were used to study their effectiveness in detecting microcalcifications using a database of 39 mammograms containing 41 microcalcifications. It was reported that the db20 had the best performance in detecting clustered microcalcifications. We will use this result in our study. In [15], image enhancement techniques were applied to a set of 100 mammograms (58 with calcification clusters). The results reported show that wavelet-enhanced digital mammograms may assist radiologists in diagnosing calcifications directly from computer monitors and may compensate for current technologic limitations. Eckstein et al in [5] studied human visual psychophysics in the context of mammographic images and the dependence of human performance in the diagnosis process on the type of noise that exist in the mammogram. Yoshida et al in [36,37] used the matching pursuit to detect microcalcifications in mammograms. The data base they used consisted of 82 region of interest (ROIs), half of which contained microcalcifications and half selected randomly

from normal areas of mammographic images. Using orthogonal wavelets with fixed weights, he reports a sensitivity of 82% and specificity of 75% in detecting microcalcifications and with optimally weighted wavelet packets he reports an improvement in the sensitivity from 82% to 92%.

Wu et al in [34] developed a convolution neural network that is used to classify benign and malignant microcalcifications in radiograph of pathologic specimens (biopsy specimens). In biopsy specimens, scatter radiation recorded on films is reduced because there is less underlying tissue around microcalcifications than in normal mammograms. As a result, microcalcifications in radiographs of biopsy specimens are generally more clearly represented than those in regular mammograms[34]. With this database he reports the average results of 80% for sensitivity and 85% specificity (compared to the average performance of radiologists of 80% sensitivity and 20% specificity).

It is important to note that the mammograms we used in this study are from an early stage and so they are harder to classify. At a later stage, the features distinguishing malignant from benign develop and make classification easier.

1.10 Overview of the Thesis

This thesis will focus on two applications which make use of wavelet analysis: the wavelet based method in image query and the classification of mammographic images. In the first application we suggest the integration of the Karhunen Loeve transform (Hotelling transform) in the wavelet based approach in image query to improve on the sensitivity of the method to displacement and rotation of the query image. In the second application we apply KLT and wavelet packet analysis to derive feature vectors based on the Joint best basis and a variant of the approximate KLT developed by Wickerhauser. We experiment with a few variants of the Joint best basis and provide the classification results using as classifiers the k-nearest neighborhood (k-nn) and Fisher's linear discriminant.

Chapter 2 provides the mathematical preliminaries of the topics used in this thesis. We start with a basic theory of wavelets presenting some aspects of wavelet analysis: the multiresolution properties (the

relation between wavelet analysis and nested linear spaces), the construction of the scaling and wavelet functions, the fast wavelet transform and the decomposition and reconstruction processes. We will describe both the 1-dimensional and 2-dimensional wavelet analysis and proceed to describe the relation of wavelet analysis to filter banks (the signal processing aspect of wavelet analysis). We then present the material relevant directly for classification, the importance of feature vectors, the relation of feature vectors and the transforms we use in this study and the two measures of the energy compaction we use, the *transform coding gain* and the *accumulated variance*. This will be followed by a description of wavelet packet analysis which is a generalization of wavelet analysis and provides a finer partition of the high frequency portion of the spectrum of the signal, the best basis and the Joint best basis which is an extension of the best basis to a population of vectors. We also provide a mathematical review of principal component analysis (the Karhunen Loeve transform), its application in reducing the representation of a signal to a smaller number of coordinates and its application in the geometric manipulation of images. The approximate KLT will then be presented which composes KLT and the Joint best basis. We end chapter 2 with a description of Fisher's LDA which is one of the classifiers we use.

In chapter 3 we describe the methodology we use in this study to apply the KLT and wavelet packet analysis to our mammographic images. This includes the application of the Hotelling transform to align a query image in some reference coordinates to improve errors due to misalignment of the query image relative to the target image. We present a few image enhancement techniques, including local image normalization and adaptive filtering to enhance features of interest and remove background structure. We present the derivation of shift invariant statistics for our mammographic images by representing each mammographic image by a large ensemble of overlapping segments sampled from the image. We then present the various feature vectors used for classification in conjunction with the k-nn classifier and Fisher's Linear Discriminant.

Chapter 4 will include the experimental results of this study. We describe the experiments starting with the poorest classification performance and ending with the best classification performance. We begin with the effects of the adaptive image enhancement in different forms and its importance to our study.

We continue by comparing the various bases in terms of their transform coding gain and the accumulated variance. We compare distribution of variance in the original basis, the standard wavelet basis, the KL basis, the Joint best basis and the approximate KL basis and analyze the results in view of the theory presented in chapter 3. We then provide classification results using various feature vectors based on the Joint best basis and the approximate KL basis.

Section 5 provide a summary of the results and discussion on some future research along the lines of this study.

Chapter 2

Mathematical Preliminaries

In this chapter we present a basic mathematical frame of wavelet theory and its applications as well as the Karhunen Loeve Transform (KLT) and related topics used in this study. Most of the material on wavelet theory is based on references [4,10,12,14,21,27,29,30,3]; the material relating to the application of wavelet analysis is based on references [17,25,33]. Some of the illustrating examples are based on examples from [17,33] and some of the related concepts which will be described, such as applications of KLT, Fisher's LDA, transform coding gain, etc. are based on [7,9,13,26].

We start with a general theory of wavelet analysis in the context of multiresolution and nested linear spaces. We show the relation between the wavelet transform and filter bank, a relation that emphasizes the signal processing aspect of wavelet analysis. We present the continuous and discrete wavelet transforms, the 1-D and 2-D fast wavelet transform. We continue to describe the wavelet packet analysis which provides a finer partition of the frequency space. We then present the best basis algorithm developed by R. Coifman and M. Wickerhauser which selects a basis with some useful properties from the wavelet packet table of a signal. This will be followed by the Joint best basis, which is an extension of the best basis applied to an ensemble of signals. We will define the concepts of information cost function which will be used to compare the representation of a signal in different bases and entropy, which is one form of information cost function. This will be followed by a formal presentation of the Karhunen Loeve Transform (KLT) known also as Principal Component Analysis (PCA) and its application in geometric manipulation of images. We then introduce the Approximate Karhunen Loeve Transform, developed by Wickerhauser which comprises the KLT and Joint best basis and which will be an important basis in this study. We will provide also a brief description of Fisher Linear Discriminant Analysis and the K-Nearest Neighborhood (k-nn) which will serve as classifiers in this study.

2.1 A Brief Review of Wavelet Theory

In this section we provide a brief review of wavelet theory from a few different perspectives: wavelet and linear subspaces, wavelet analysis as an orthogonal transform with high transform coding gain and the relation between wavelet analysis and filter banks.

2.1.1 Linear Space and Wavelets

We present the mathematical frame that reveals the multiresolution aspect of the wavelet transformation. We start with the definition of a set of subspaces, V_n , $n \in \mathbb{Z}$, with certain properties.

Every subspace V_n , $n \in \mathbb{Z}$, satisfies the following axioms:

1. Scaling:

$$f(x) \in V_n \iff f(2x) \in V_{2n}$$

2. Inclusion:

$$V_n \subset V_{n+1} \text{ for all } n$$

3. Density:

$$\text{Closure} : \left\{ \bigcup V_n \right\} = L^2(\mathbb{R})$$

4. Maximality:

$$\bigcap V_n = \{0\}$$

5. Basis: $\exists \Phi(x)$ such that the set $\{\Phi(x - k)\}_{k \in \mathbb{Z}}$ is an orthonormal basis for V_0

i.e.

$$f(x) \in V_0 \implies f(x) = \sum_k a_k \Phi(x - k)$$

Now, by axiom (5) the subspace $[\{\Phi(x - k)\}_{k \in \mathbb{Z}}]$ is an orthonormal basis for V_0

and so by axiom (1):

$2^{n/2} \{\Phi(2^n x - k)\}_{k \in \mathbb{Z}}$ is an orthonormal basis for V_n .

The union of all subspaces,

$$\bigcup_{n \in \mathbb{Z}} 2^{n/2} \{\Phi(2^n x - k)\}_{k \in \mathbb{Z}}$$

is a basis for $L^2(\mathbb{R})$ but it is not an orthogonal basis. Since $V_n \subset V_{n+1}$, the subspaces are not mutually orthogonal.

2.1.1.1

Creating an orthonormal basis by defining a wavelet subspace

Based on the properties of the subspaces we have:

$$V_1 = V_0 \oplus W_0$$

$$V_2 = V_1 \oplus W_1 = V_0 \oplus W_0 \oplus W_1$$

.

.

.

$$V_{n+1} = V_0 \oplus W_0 \oplus W_1 \oplus \dots \oplus W_n$$

and the space of measurable functions can then be expressed as:

$$L^2(\mathbb{R}) = V_0 \oplus W_0 \oplus W_1 \oplus \dots$$

We now can express $L^2(\mathbb{R})$ in terms of the wavelet subspaces only.

To see that, we observe that

$$V_0 = V_{-1} \oplus W_{-1} = V_{-2} \oplus W_{-2} \oplus W_{-1} = \dots \oplus W_n, \quad n \in \mathbb{Z}$$

and so $L^2(\mathbb{R})$ is an orthogonal sum of the wavelet subspaces:

$$L^2(\mathbb{R}) = \bigoplus_{n \in \mathbb{Z}} W_n$$

Example: The Haar Wavelet

The Scaling function, $\Phi(x)$ and the wavelet function are defined:

$$\Phi(x) = \begin{bmatrix} 1 & \text{for } 0 < x \leq 1 \\ 0 & \text{otherwise} \end{bmatrix}$$

$$\Psi(x) = \begin{bmatrix} 1 & \text{for } 0 < x \leq 1/2 \\ -1 & \text{for } 1/2 < x \leq 1 \end{bmatrix}$$

The two functions satisfy:

$$\begin{aligned} \int \Phi(x) dx &= 1 \\ \int |\Phi(x)|^2 dx &= 1 \\ \int \Psi(x) dx &= 0 \\ \int |\Psi(x)|^2 dx &= 1 \end{aligned}$$

The union of scaled shifted version of the wavelet function

$$_{n \in \mathbb{Z}} \bigcup 2^{n/2} \{ \Psi(2^n x - k) \}_{k \in \mathbb{Z}}$$

provides an orthonormal basis(the Haar basis) for $L^2(\mathbb{R})$.

(Note: the factor $2^{n/2}$ preserves the L^2 norm).

Now, given a function $f(x)$, what would be its expansion using the Haar basis? Since the basis is orthonormal we can find the expansion coefficients easily.

$$\begin{aligned} f(x) &= \sum_{n, k \in \mathbb{Z}} \beta_k^n 2^{n/2} \Psi(2^n x - k) \\ \beta_k^n &= \int f(x) 2^{n/2} \Psi(2^n x - k) dx \end{aligned}$$

Where β_k^n are the wavelet coefficients, n represents the scale and k the displacement of the wavelet function.

2.1.2 The Dilation Equation and the Construction of the Scaling and Wavelet functions

First we show how the scaling function is constructed using the dilation equation. The basic dilation equation is a two-scale difference equation[30]:

$$\Phi(x) = \sum_k c_k \Phi(2x - k)$$

where c_k represent the lowpass filter coefficients. To find a solution for $\Phi(x)$ we first require a normalized solution that satisfies:

$$\int \Phi(x)dx = 1$$

Integrating both sides of the equation for $\Phi(x)$, it follows that:

$$\sum c_k = 2$$

For example, the 'Haar' scaling function is constructed from the filter coefficients:

$$c = (1, 1)$$

which yield the box function:

$$\Phi(x) = 1 \text{ for } 0 < x < 1$$

and its scaled version:

$$\Phi(2x) = \left\{ \begin{array}{ll} 1 & \text{for } 0 < x < 1/2 \\ 0 & \text{for } 1/2 \leq x < 1 \end{array} \right\}$$

and $\Phi(2x - 1)$ is a translated scaled version of $\Phi(x)$. Using the dilation equation we can continue to derive $\Phi(4x), \Phi(8x), \dots$

The wavelet function then can be derived from the equation:

$$W(x) = \sum (-1)^k c_{N-k} \Phi(2x - k)$$

where N is the number of coefficients in c . Note that $(-1)^k c_{N-k}$ represents the coefficients of the complementary high pass filter in the filter bank. Applying this equation for the 'Haar' case with $c = (1, 1)$, we get the Haar wavelet:

$$W(x) = \Phi(2x) - \Phi(2x - 1)$$

2.1.3 The Fast Wavelet Transform

In general given a function $f(t)$, its *Continuous Wavelet Transform* is defined to be:

$$C(scale, position) = \int f(t) \Psi(scale, position, t) dt$$

where $\Psi(scale, position)$ is a scaled and shifted version of the wavelet function used. The result of the transform are the coefficients C , functions of scale and position. This transform is continuous with respect to scale and position and so involves an enormous amount of calculation and generates a lot of wavelet coefficients. It was found that if we restrict the analysis only to certain scales and positions, we can still get a set of coefficients which can be used to reconstruct the original signal without compromising accuracy. In the Discrete Wavelet Transform (DWT), we use dyadic scales and positions (based on powers of 2). An efficient and fast decomposition and reconstruction algorithm was developed in 1988 by Mallat (like the FFT for the Fourier transform) which uses a set of filters and executes the wavelet transform in $O(N)$ steps.

2.1.4 One-level and multilevel decomposition of a signal

The scheme for one stage decomposition involves applying low pass and high pass filters to the signal, each followed by a downsampling (decimation) step so that the total number of coefficients equals the length of the original signal. The output of the low pass filter provides the approximation coefficients and the output of the high pass filter provides the details. In multilevel decomposition of a signal, the above process is iterated on the approximation coefficients with successive approximations being decomposed in turn. This process can be continued until the approximation coefficients consist of a single value. In practice the number of levels is chosen based on the signal or a criterion such as entropy. The final result of this process is the *wavelet decomposition tree* which enables the reconstruction of the signal at different levels of resolution.

2.1.5 The Reconstruction Process

The approximation and detail coefficients of a signal can be used to reconstruct the original signal. This is called the *Inverse Discrete Wavelet Transform (IDWT)*. The synthesis is a reverse process. One stage of synthesis is composed of applying an upsampling and filtering to the approximation and detail

coefficients of the last stage of the wavelet analysis. This would result in the approximation of the signal at the next level. We can continue this process by upsampling and filtering the new approximation coefficients and the next set of detail coefficients to produce the approximation of the signal at the next level. If we continue this process, the final synthesis step would produce the original signal (the process can be iterated $\log_2(N)$ times where N is the length of the signal). More precisely, if the analysis and synthesis filters satisfy certain conditions (QMF) they will provide perfect reconstruction. Note that at each stage of the reconstruction process, we get an approximation of the signal at a different level of resolution.

The discrete version uses a two channel subband coder implemented via a set of quadrature mirror filters (QMF). The derivation of the four filters used for the decomposition and reconstruction begins with the dilation equation and will be described later.

2.1.6 Wavelets and Filter Banks

Wavelets and filter banks are closely related and this relationship provides the link between wavelet analysis and signal processing. The choice of the filters determines the shape of the wavelet and whether it provides a perfect reconstruction or not. In constructing a wavelet, the first step is to derive the filter bank which consists of an analysis stage H_0, H_1 and a synthesis stage F_0, F_1 [30]. Quadrature mirror filters (QMF) have the following property: the high pass filter is a mirror image of the low pass (in squared magnitude) with respect to the middle frequency $\pi/2$ (called the quadrature frequency). In the analysis stage the signal is analyzed by a convolution-subsampling step using the low pass and high pass filters. The synthesis step consists of upsampling and filtering.

We will follow the notation used in [25] to describe the relation of wavelet analysis and filter banks for discrete signals.

Let $f = \{f_k\}_{k=0}^{2K-1}$ be a real valued vector of even length $2K$ representing the signal to be analyzed. Let $h = \{h_k\}_{k=0}^{L-1}$ and $g = \{g_k\}_{k=0}^{L-1}$ be the impulse response coefficients of the low pass and high pass filters, respectively.

The analysis step of the signal f consists of the convolution-subsampling operations represented by the operators H and G :

$$(Hf)_k = \sum_{l=0}^{L-1} h_l f_{2k+l}$$

$$(Gf)_k = \sum_{l=0}^{L-1} g_l f_{2k+l}$$

for $k = 0, 1, \dots, K-1$. Let $c = \{c_k\}_{k=0}^{K-1}$ and $d = \{d_k\}_{k=0}^{K-1}$ be the coefficients resulting from the analysis stage H and G respectively.

The synthesis of the signal from these coefficients is done by H^* and G^* , the adjoint operators of H and G defined by the upsampling-convolution operations:

$$(H^*c)_k = \sum_{0 \leq k-2l < L} h_{k-2l} c_l$$

$$(G^*d)_k = \sum_{0 \leq k-2l < L} g_{k-2l} d_l$$

for $k=0,1,\dots,2K-1$. The two sets of filters, the analysis filters H, G and the synthesis filters H^*, G^* are called quadrature mirror filters (QMF) if they satisfy the orthogonality conditions:

$$HG^* = GH^* = 0$$

$$HH^* + G^*G = I$$

where I is the identity operator. The orthogonality conditions are equivalent to perfect reconstruction.

Given a set of low pass and high pass filters, h and g , what would be the restrictions on these filters so that the operators H, G are orthogonal (and provide perfect reconstruction)? To answer this question we define

$$m_0(\xi) = \sum_{k=0}^{L-1} h_k e^{ik\xi}$$

$$m_1(\xi) = \sum_{k=0}^{L-1} g_k e^{ik\xi}$$

then H, G are QMF if and only if the following matrix is unitary:

$$\begin{pmatrix} m_0(\xi) & m_0(\xi+\pi) \\ m_1(\xi) & m_1(\xi+\pi) \end{pmatrix}$$

The relation between the low pass and high pass filters g, h is given by:

$$g_k = (-1)^k h_{L-1-k}$$

i.e. we reverse the order of the coefficients in h and change the sign of the odd numbered coefficients.

As the frequency response of the low pass and high pass filters overlap, this will result in aliasing in each channel; therefore, the synthesis filters, F_0, F_1 , have to be adapted to the analysis filters to cancel the aliasing. Filter banks which provide a perfect reconstruction are called *biorthogonal* filter banks. The synthesis bank is the inverse (in terms of the matrix operations) of the analysis bank.

Next we illustrate the relation between wavelets and filter banks using the MATLAB wavelet toolbox. We show the derivation of the 4 tap Daubechies wavelet (db2) from the low pass and high pass filter bank associated with it [30, 17].

The low pass filter coefficients (impulse response) corresponding to db2 are:

$$c = (1 + \sqrt{3}, 3 + \sqrt{3}, 3 - \sqrt{3}, 1 - \sqrt{3}) / (4\sqrt{2})$$

This low pass filter has double zeros at the Nyquist frequency (the highest frequency) and the set of coefficients satisfy the following orthonormal properties:

The sum of the coefficients is 2.

The sum of the square of the coefficients is 1 (euclidean norm is 1);

The vector is orthogonal to its double shift, i.e. $\sum c(k) * c(k-2) = 0$.

The high pass filter coefficients, $d(k)$, corresponding to db2 can be derived from:

$$d(k) = qmf(c(k)) = (-1)^k c(N-k) \quad N = 4, \quad k = 0, 1, 2, 3$$

This gives:

$$d = (1 - \sqrt{3}, -(3 - \sqrt{3}), 3 + \sqrt{3}, -(1 + \sqrt{3})) / (4\sqrt{2})$$

This is the impulse response of the high pass filter and it satisfies the following properties:

The sum of the coefficients is 0.

The sum of the square of the coefficients is 1 (euclidean norm is 1);

The vector is orthogonal to its double shift, i.e. $\sum d(k) * d(k - 2) = 0$.

We use these two filters to derive the two sets of decomposition and reconstruction filters:

$$LowPassReconFilter = c(k)$$

$$HighPassReconFilter = d(k)$$

$$LowPassDecompFilter = reverse(c(k))$$

$$HighPassDecompFilter = reverse(d(k))$$

Fig. 2.1 provides the graphs of the four filter coefficients and their frequency for the db6, 12 tap wavelet. They were derived similar to the derivation of the QMF based on db2.

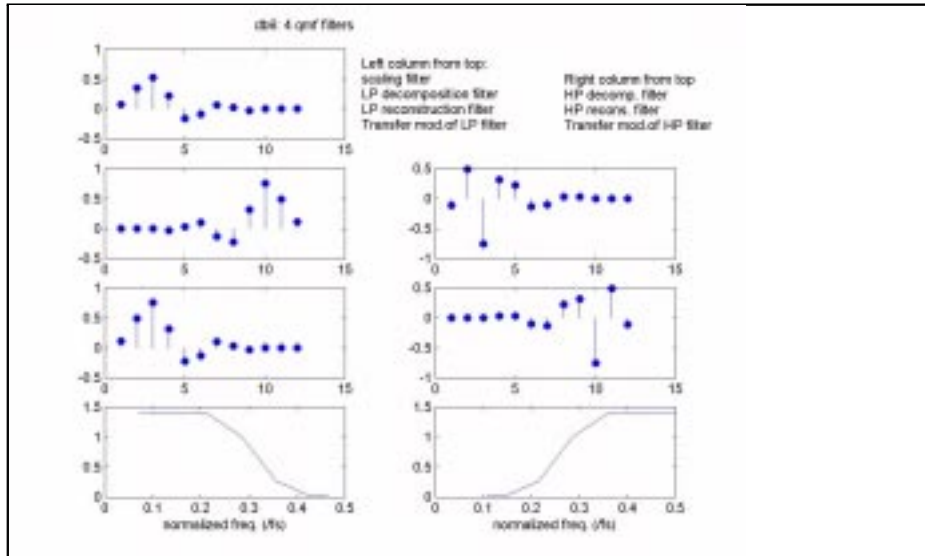


Figure 2.1. db6 4 qmf filters. Left panel from top: Scaling filter, LP decomposition filter, LP reconstruction filter, transfer modulus of LP filter. Right column from top: HP decomposition filter, HP reconstruction filter, transfer modulus of HP filter.

2.1.7 Vanishing Moments of a Wavelet Function

One of the properties of a wavelet function is the number of its vanishing moments. This property relates to the compression rate achieved using the wavelet.

For example, the Haar wavelet has 1 vanishing moment (the coefficients corresponding to intervals where the analyzed function is constant will be zero), db2 has two vanishing moments (the coefficients corresponding to intervals where the function is either constant or linear will be zero) and db3 has 3 vanishing moments (the coefficients corresponding to intervals where the function is either constant, linear or a polynomial of degree 2, will be zero) and so on. In general, the higher the number of the vanishing moments, the higher the compression rate is. Of course the higher the vanishing moment, the wider is the support of the filter (larger number of coefficients). For a specific signal, it is possible to find an upper bound on the detail coefficients using a Taylor series expansion of the analyzed signal[29].

2.1.8 1-D Wavelet Analysis

We next illustrate some of the capabilities of 1-dimensional wavelet analysis. The illustrations are based on tutorial examples that appear in [17].

The first example illustrates how wavelet analysis can provide local information on a signal and detect a discontinuity, even a minute one. We analyze a signal that is composed of two exponentials connected together. The signal is perfectly smooth except at the point where the two exponentials meet, but the discontinuity is not visible. In fig 2.2 we apply a two level wavelet analysis using db4 and zoom on the detail coefficients.

Because the signal is perfectly smooth we need a wavelet function with a high level of regularity (high number of vanishing moments) to detect the singularity. The discontinuity is clear in the detail coefficients as can be seen in the picture. Fourier analysis could not provide information on this 'slight' discontinuity.

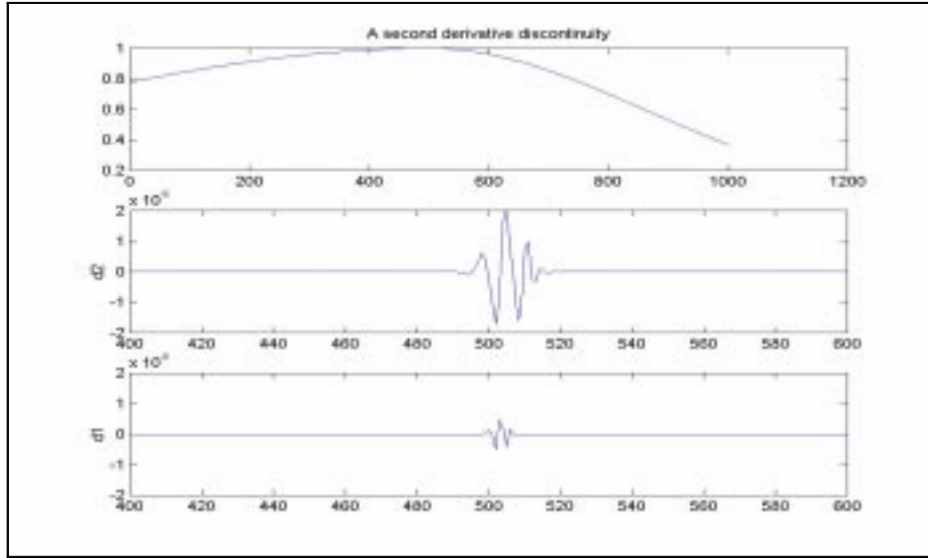


Figure 2.2. A signal with second derivative discontinuity. Top panel- signal, mid panel-second detail coeff's, bottom panel-first level detail coeff's.

In the next example we illustrate the distribution of white noise at different levels of the wavelet analysis. The signal we analyze is pure white noise (the energy is distributed evenly over all frequencies). We apply a wavelet analysis at level 2, using db3 and examine the energy distribution at different levels of the wavelet decomposition. In general, the details at the first level of decomposition contain the highest frequency components, the details at the second level contain the next layer of high frequencies and so on. As can be seen in fig. 2.3, the details at all levels do not have any regularity and look like noise-type signals. This is due to the fact that the signal is composed of white noise including all frequencies. But as we go deeper in the decomposition level, color is introduced in the coefficients (since at each level, a certain level of frequencies is removed). Also the variance (of the amplitude) of the details decreases as the level of decomposition increases (the variance decreases two fold from one level to the next). This observation is important in de-noising a signal.

Next is an example from [17] (p.6-83) which illustrates the de-noising capabilities of wavelet analysis, using soft thresholding that is adapted to different levels of the detail coefficients (and so it is not

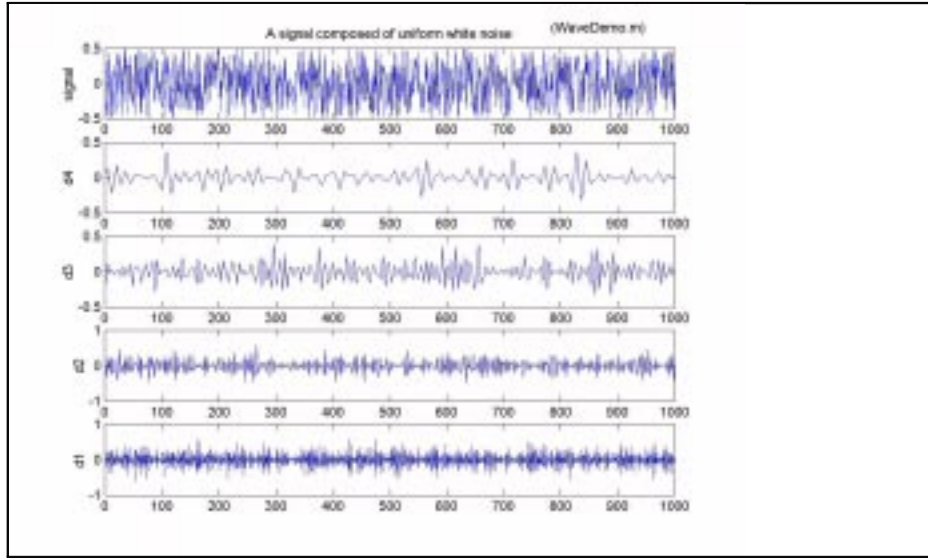


Figure 2.3. Wavelet analysis of a signal composed of white noise. From top to bottom panel: analyzed signal, detail coeff's at level 4, 3,2,1 respectively.

global filtering). The signal to be analyzed is a noisy electrical signal (due to random fluctuations in electrical consumption). Fig. 2.4 provides the signal and the de-noised version.

2.1.9 2-D Wavelet Analysis

There are two schemes for 2 dimensional wavelet analysis: The *standard wavelet transform* and the *nonstandard wavelet transform* [29]. In the standard wavelet transform, we first apply a full one dimensional transform to each row of the image, then we apply a full one dimensional transform to each column of the result of the first step. The first step provides the average and details for each row. The result of the second step would be an overall average and the rest are details of the transformation.

The second method, used by the MATLAB Wavelet Toolbox, is the nonstandard wavelet transform. In this scheme we apply the following steps:

1. One step of transformation to each row
2. One step of transformation to each of the columns in the result

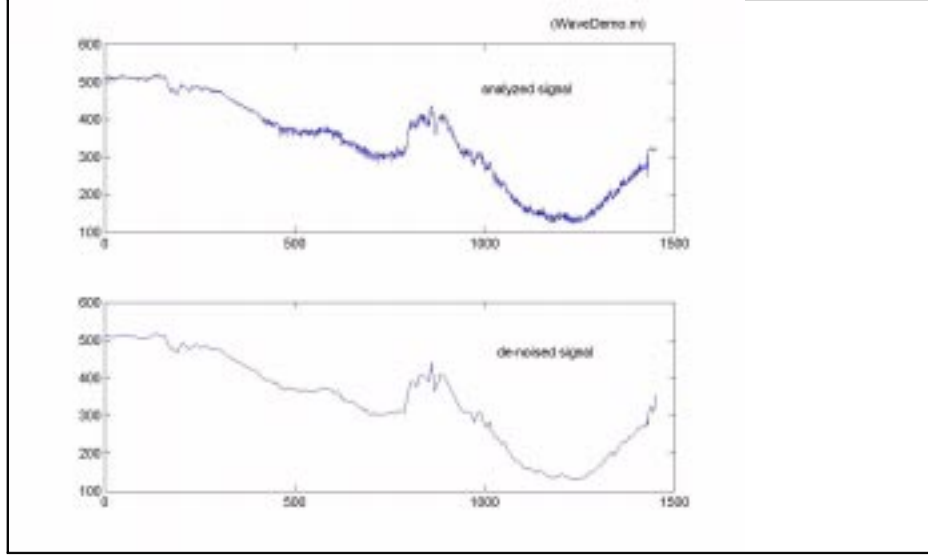


Figure 2.4. De-noising of a signal. Top panel- noisy electrical signal, bottom- de-noised version.

These two steps result in a quadrant of average and three quadrants of details: horizontal, vertical and diagonal.

3. We repeat steps 1,2 on the quadrant of the average.

The final result of this scheme is an overall average of the image and layers (their number depends on the number of levels of decomposition) consisting of quadrants of details: horizontal, vertical and diagonal. As in the case of 1-d wavelet analysis, this structure enables reconstruction of the image at different levels of resolution.

The basis functions for the 2-D wavelet transform are derived by a tensor product of the one dimensional scaling and wavelet functions.

If $\Phi(x)$ is the one dimensional scaling function and $\Psi(x)$ is the one dimensional wavelet function, we define the following 2-dimensional scaling and wavelet functions: $\Phi\Phi(x, y)$, the two-dimensional scaling function

$$\Phi\Phi(x, y) = \Phi(x)\Phi(y)$$

and three wavelet functions: $\Phi\Psi(x, y)$, $\Psi\Phi(x, y)$, and $\Psi\Psi(x, y)$, as follows:

$$\Phi\Psi(x, y) = \Phi(x)\Psi(y)$$

$$\Psi\Phi(x, y) = \Psi(x)\Phi(y)$$

$$\Psi\Psi(x, y) = \Psi(x)\Psi(y)$$

Using these wavelet functions we can create a set of scaled and translated versions of the three wavelet functions. The nonstandard basis will consist of the single scaling function and a set of scaled and translated versions of the three wavelet functions[29].

2.1.10 2-D wavelet analysis of images

We illustrate the nonstandard decomposition of 2-d images using the standard image 'woman'. We will apply one level of decomposition to the image resulting in the approximate coefficients and the horizontal, vertical and diagonal details. Then we will apply another level of decomposition to the approximation coefficients to get a second set of approximation coefficients and horizontal, vertical and diagonal coefficients. We will 'arrange' the first set of three detail coefficients and the second set of three detail coefficients and approximation in a single matrix which is displayed in fig.2.5. Note that the smaller replica of the image is of size 1/4x1/4 compared to the original image. Having three sets of details can be very useful for certain applications. For example we may be interested in some feature which is characterized by certain frequencies aligned diagonally in the image. The coefficients corresponding to these features will appear in the diagonal quadrants.

2.1.11 De-noising of 2-dimensional images

As in the case of a 1-d signal, the wavelet transform can be used for de-noising of 2-d images. We illustrate the adaptive de-noising capabilities of wavelet analysis using an example from [17] (p.6-84). We use a synthetic signal to which we add normal gaussian white noise; then a threshold is computed to be used for de-noising. Fig. 2.6 shows the original image, the noisy and the de-noised signal.

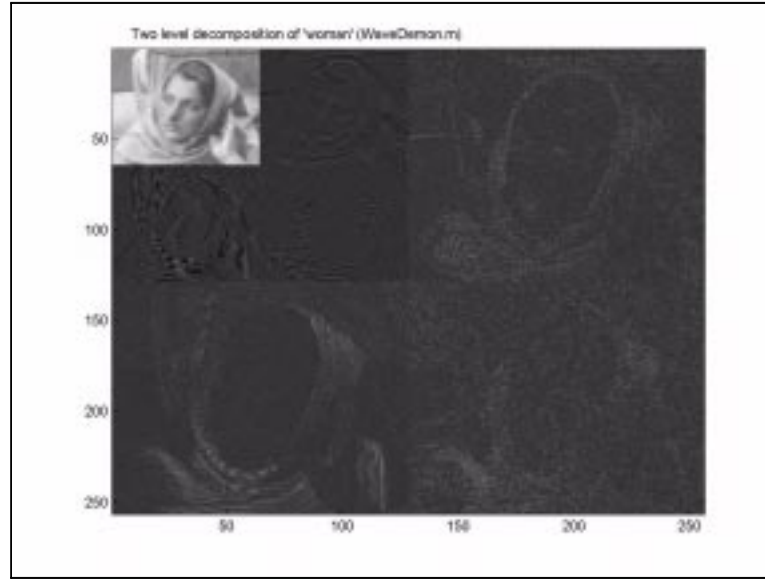


Figure 2.5. Two levels of nonstandard decomposition of the image 'woman'

2.2 Wavelet Packets

In standard wavelet analysis, one level of the analysis consists of decomposing the signal into approximation coefficients and details. In the second level of decomposition, the approximation coefficients are decomposed again into approximation and detail coefficients. Wavelet analysis does not treat all frequencies equally. It provides a finer partition of the lower half of the frequency spectrum. For many signals this may be adequate as most of the energy of the signal may be concentrated in the lower half of the spectrum. For other signals, a different approach provided by wavelet packet analysis is more adequate. Wavelet packet analysis is just a generalization of the wavelet analysis. It 'treats' all frequencies equally. In the first level of decomposition, the signal is decomposed into approximation coefficients and detail coefficients. The process then is applied to both the approximate and detail coefficients. Thus wavelet packet analysis provides a finer partition of both the high and low frequency ranges. This may be important in cases where the interesting features of the signal are hidden in the high frequency components and a finer resolution of frequency information can help extract those features. While wavelet analysis has a computation complexity of $O(n)$, wavelet packet analysis has a complexity of $O(n \log n)$ [25].

$$\begin{array}{cccccc}
ss_0 & ss_1 & ds_2 & ds_3 & sd_4 & sd_5 & dd_6 & dd_7 \\
H \swarrow \searrow & G & H \swarrow \searrow & G & H \swarrow \searrow & G & H \swarrow \searrow & G \\
sss & dss & sds & dds & ssd & dsd & sdd & ddd
\end{array}$$

Note that each row of coefficients is completely determined by the coefficients of the previous row and the two convolution-down sampling filters H, G . Also, each row of coefficients can be derived using the adjoint operators of H, G which we denote by H^*, G^* . These adjoint operators are upsampling-convolution operators and they enable the reconstruction of the whole packet coefficients from the coefficients at the leaves. The data structure used by MATLAB [17] does not provide all the coefficients of a wavelet packet corresponding to a signal. Rather, it holds only the coefficients at the leaves and, using the two adjoint operators, it can compute the coefficients at any level and position within a level. Wave-lab [32] does have commands that provide all the coefficients in a packet. The wavelet packet of a signal is used to derive its best basis, and the wavelet packets of an ensemble of vectors are used to derive the Joint best basis as we describe later.

2.2.1 Orthonormal Bases and Information Cost Functions

An important property of wavelet packets (32) is the fact that they contain many orthonormal bases. This is due to the properties of the QMF applied in the derivation of the packet. Any subset of these coefficients with a 'horizontal projection' that has complete coverage and such that the projection of no two coefficients overlap provides an orthonormal basis. The first row (the original data) represents the coefficients in the euclidean basis and any other row provide an orthonormal basis. The coefficients in the standard wavelet basis are: $sss, dss, ds_2, ds_3, d_4, d_5, d_6, d_7$.

The wavelet packet provides many bases which can be used to represent the signal. To compare the different representations, we introduce the notion of information cost that has its origins in informa-

tion theory. The various information cost functions provide a measure of how evenly the 'energy' of a signal (vector) is distributed among the various coordinates. A 'good' representation for the purpose of compression, for example, would have a few large values and all the others would be very small. The information cost function we choose has to be additive, as this is a necessary property in the best basis algorithm that will be described later.

Formally, given a sequence $\{s\} = (s_1, s_2, \dots)$, and μ , a real valued function defined on $[0, \infty)$, the real valued functional M is said to be additive if

$$\begin{aligned} M(s) &= \sum_i \mu(|s_i|) \\ \text{and } \mu(0) &= 0 \end{aligned}$$

There are various information cost functions which can be found in [17],[33]. We will use an information cost functional based on Shannon entropy defined to be:

$$H(s) = - \sum_i s_i^2 \log(s_i^2)$$

We also assume the sequence is normalized, i.e. $\sum_i s_i^2 = 1$.

2.2.2 The Best Basis Algorithm

As we have seen, the wavelet packet of a signal provides many bases that can be used to represent a signal. We would like to search for the basis with the lowest information cost. A fast and efficient algorithm that has a complexity of $O(n[\log n])$ was developed by Coifman and Wickerhauser. We present the algorithm following the notations in [25] and [33]. The algorithm will find the basis that minimizes a given information cost as long as it is additive. Note that given a signal, one can derive more than one wavelet packet for the signal using different QMFs. Each will provide a library that will have a best basis. Having the best basis of each library, one can choose the basis among these bases using again an information cost such as Shannon entropy.

Given a finite sequence $\{s\} = (s_1, s_2, \dots, s_N)$, which may represent the coefficients of a signal in euclidean coordinates, we first compute its wavelet packet and arrange it in a binary tree. Each entry in the table, $w_{j,n} = (w_{j,n,0}, w_{j,n,1}, w_{j,n,2}, \dots, w_{j,n,2^j-1})$ is a vector called a crystal, containing a set of coefficients called atoms. The atom $w_{j,n,k}$ represents the coefficient at scale j , frequency block n , and position k within frequency block n . Atoms corresponding to a lower level of decomposition (lower entries in the table), provide more localized spatial/frequency information on the signal. Viewing the wavelet packet table in a tree form for an orthogonal wavelet, each complete subtree is a basis for the signal, and so for a vector $\{s\}$, decomposed at level L , there are 2^L complete binary subtrees, each being a basis (the wavelet transform is one of these bases). The best basis algorithm searches for the complete subtree which minimizes some additive cost functional. As the wavelet packet coefficients are arranged in a tree structure, an efficient search algorithm is possible for finding the best basis with computational complexity of $N \log N$. For images, the wavelet packet analysis is composed of a quaternary tree containing the approximation coefficients and the horizontal, vertical and diagonal details.

The best basis algorithm utilizes the fact that the information cost is additive and, starting from the leaves (bottom) of the binary tree, it 'prunes' the binary tree by comparing the entropy of the parent node to the sum of the entropies of its two children nodes and chooses the representation with the lower information cost.

The Best Basis Algorithm:

1. Specify the QMF to be used to derive the wavelet packet, the maximum depth of wavelet packet decomposition J and an information cost functional.
2. Derive the wavelet packet table coefficients using the QMF chosen at level J , and arrange it in a tree form (i.e. each node represents a crystal in the packet table).
3. Compute the entropy (information cost) of each node in the binary tree of the wavelet packet.
4. Mark all the leaves.

5. Starting from the nodes at the bottom of the binary tree (leaves), compare the entropy of a parent node to the sum of entropies of its two children nodes. If a parent node has a lower entropy value, we mark the parent, otherwise we do not mark the parent but replace its entropy with the sum of the entropies of its two children.

6. Since the binary tree is of finite size, step 5 will be completed in a finite amount of time and we reach the root (the original signal).

7. Starting from the top we select the topmost marked nodes which represent the coefficients of the signal in the best basis relative to the cost functional and the QMF which we selected.

We illustrate the algorithm with an example taken from [33]. Fig 2.7 shows the binary tree of entropy values of a signal with 3 levels of wavelet packet decomposition. The numbers represent the entropy values of the wavelet packet coefficients for each crystal. (As was described before, each atom represents a coefficient corresponding to a scale, frequency and position within the frequency block. Crystals represent a set of atoms corresponding to certain scale and frequency.) In the first step we mark all the leaves (marked with an asterisk in fig 2.7).

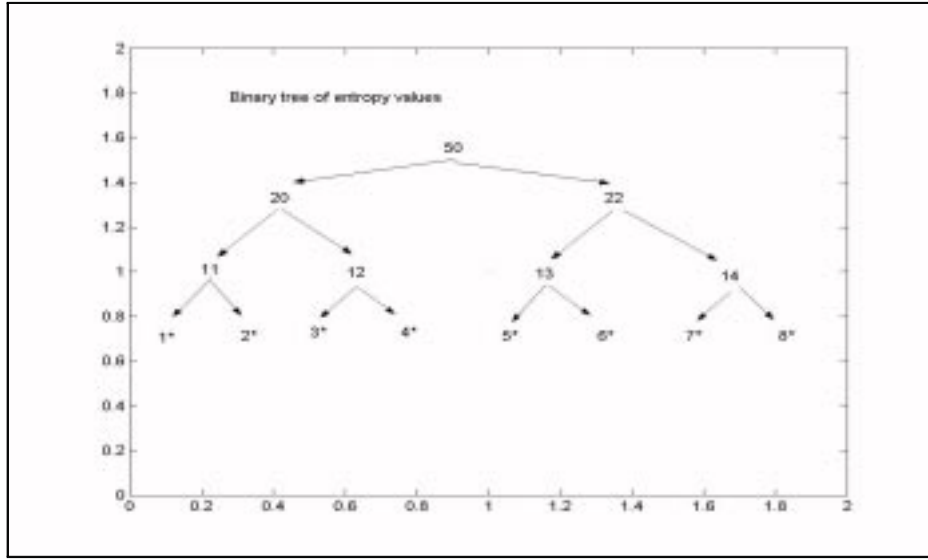


Figure 2.7. First step in best basis search: Mark all bottom nodes in the binary tree representing the entropy values of each crystal(represented by a node)

In fig.2.8, we perform step 5 in the algorithm, comparing the entropy of the parent at each node to the sum of its two children and replacing the entropy value of the father node if that sum is lower. For example, the sum of the entropies of the children of the most left bottom node is $1+2=3$, while the entropy of the parent node is 11, so we replace 11 by 3, indicated by 3(11) and we do not mark the parent node. On the other hand, the most right bottom node has a parent with entropy value of 14, lower than the sum of its two children: $7+8=15$, and so we leave the value 14 and mark the parent node. This step is carried from bottom nodes up.

In fig.2.9 we perform the last step in the best basis search. We select the topmost marked nodes (enclosed by a rectangular box). Note that the projection of the set of crystals (nodes) selected provides a full 'coverage' of the space and that there is no overlap, meaning the basis is orthogonal.

This algorithm is implemented both in MATLAB [25] and Wavelab [32], with the Wavelab package having more versatile functions that we will be using to implement the Joint best basis described next.

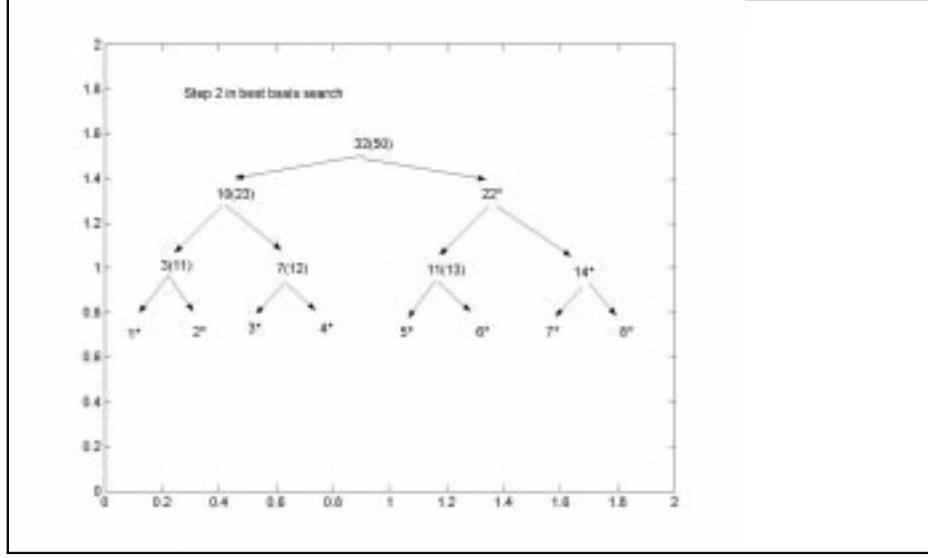


Figure 2.8. Second step in search for a best basis: compare the entropy of parent and its children and mark all nodes (marked with an asterisk) with lower cost, starting from the bottom.

2.2.3 The Joint Best Basis (JBB)

The JBB is an extension of the best basis to an ensemble of signals. Having an ensemble of vectors, we would like to find a basis which provides a good representation of all vectors, in average. With respect to a certain QMF, the JBB minimizes the information cost among all the other bases determined by the QMF.

Let $V = \{V_1, V_2, \dots, V_N\} \subset R^d$ be an ensemble of vectors which may represent an ensemble of 1-d signals or an ensemble of images. We apply wavelet packet analysis to all vectors. Following the notation in [33], we denote by $\lambda_{sf}^{(n)}(p)$, the wavelet packet coefficient corresponding to V_n at scale s , frequency block f and index(position) p within block f . The standard deviation of this coefficient (computed over all the vectors in the ensemble) is:

$$\sigma_{sf}(I)(p) = \left[\frac{1}{N} \sum_{n=1}^N [\lambda_{sf}^{(n)}(p)]^2 - \left[\frac{1}{N} \sum_{n=1}^N \lambda_{sf}^{(n)}(p) \right]^2 \right]^{\frac{1}{2}}$$

Computing the JBB involves the following steps:

1. We compute the standard deviation $\sigma_{sf}(I)(p)$, for all the coefficients in the wavelet packets, resulting in a

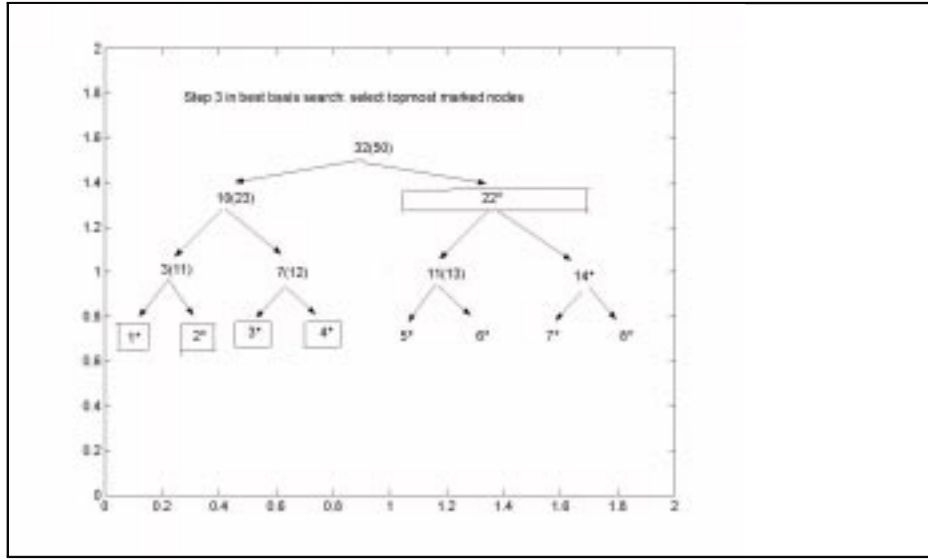


Figure 2.9. Final step in search for a best basis: select topmost marked nodes (enclosed by rectangulars) to get the best basis.

standard deviation packet table for this ensemble.

We apply the best basis algorithm to the *standard deviation packet table* (i.e. it replaces the packet table in the BB algorithm), to search for a best basis. The resulting basis is the Joint BB.

Next we use the JBB to extract a set of coefficients from the standard deviation packet table and sort them by their magnitude. The indices corresponding to the coefficients with the largest magnitude provide the coordinates which account for the largest variation of the wavelet packet coefficients of the ensemble, and these few coordinates can be used to provide the best approximation (within the library of bases determined by the QMF) to represent the ensemble of vectors.

It is worthwhile to note that as the wavelet transformation is a linear operator but the standard deviation is not, the standard deviation packet table is not a "legitimate" packet table in that the relation between coefficients in one scale to the next does not satisfy the relation determined by the QMFs.

2.3 Principal Component Analysis (Karhunen Loeve Transform)

Principal Component Analysis (PCA) known also as the Karhunen Loeve transform(KLT) or factor analysis is a linear transformation that minimizes an information cost function over all orthogonal transformations when the ensemble of vectors has a multivariate normal distribution. For an ensemble of signals with n coordinates, it has a complexity of $O(n^3)$ which makes it impractical for desktop computing when the dimension of the problem, $n > 1000$. Later we describe the *approximate Karhunen Loeve Transform*, which provides a computable basis and is the closest to KLT basis in a family of bases with the lower complexity of $O(n^2 \log n)$.

We now present the algebra of this transformation and show how it decorrelates the coordinates of the data in the original basis. Given a set k of n -dimensional real valued column vectors: X_1, X_2, \dots, X_k , the covariance matrix of the vector population is given by:

$$C_x = E[(X - m_x)(X - m_x)^T]$$

where m_x is the mean vector of the population. C_x is a $n \times n$ real symmetric matrix whose c_{ij} entry equals the covariance between the i^{th} and the j^{th} coordinates of the vector population. If $c_{ij} = 0$, then the i^{th} and the j^{th} coordinates are decorrelated. When $c_{ij} > 0$ or $c_{ij} < 0$, there is a positive or a negative correlation between the i^{th} and the j^{th} coordinates, respectively.

The Karhunen Loeve transform maps the given vector population, X , into a vector population Y (that consists of k , n -dimensional vectors) such that $m_y = 0$ and C_y , the covariance of the new vector population, is a $n \times n$ diagonal matrix; which implies that the i^{th} and the j^{th} coordinates of the new vector population are decorrelated for all $i \neq j$. To see this, let X be a set of k , n -dimensional column vectors: X_1, X_2, \dots, X_k . The mean vector of the population m_x is given by:

$$m_x = \frac{1}{K} \sum_{i=1}^K X_i$$

and the covariance matrix of the vector population X is:

$$C_x = E[(X - m_x)(X - m_x)^T] =$$

$$\begin{aligned}
&= E(XX^T) - E(Xm_x^T) - E(m_x X^T) + E(m_x m_x^T) = \\
&= E(XX^T) - m_x m_x^T - m_x m_x^T + m_x m_x^T = \\
&= \frac{1}{K} \sum_{i=1}^K X_i X_i^T - m_x m_x^T
\end{aligned}$$

Now, C_x is a $n \times n$ real and symmetric matrix; therefore it has n real eigenvalues and n real orthogonal eigenvectors[11].

Let $\lambda_1, \lambda_2, \dots, \lambda_n$ be the n eigenvalues with $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$ and V_1, V_2, \dots, V_n be the corresponding set of orthonormal eigenvectors, where V_1 corresponds to λ_1 , the largest eigenvalue. Define the matrix A whose rows are the n eigen vectors V_1, V_2, \dots, V_n :

$$A = \begin{bmatrix} V_1 & - & - & - & - & > \\ V_2 & - & - & - & - & > \\ & & & & \cdot & \\ & & & & \cdot & \\ V_n & - & - & - & - & > \end{bmatrix}$$

The Karhunen Loeve transform is then the linear transformation given by:

$$Y_i = A(X_i - m_x) \quad i = 1, 2, \dots, k$$

The mean of the vector population Y is 0 since

$$\begin{aligned}
m_y &= \frac{1}{K} \sum_{i=1}^K Y_i = \\
&= \frac{1}{K} \sum_{i=1}^K A(X_i - m_x) = \\
&= A \frac{1}{K} \sum_{i=1}^K (X_i - m_x) = 0
\end{aligned}$$

and the covariance matrix of the vector population Y is diagonal:

$$\begin{aligned}
C_y &= E[(Y - m_y)(Y - m_y)^T] = \\
&= \frac{1}{K} \sum_{i=1}^K Y_i Y_i^T - m_y m_y^T = \\
&= \frac{1}{K} \sum_{i=1}^K Y_i Y_i^T =
\end{aligned}$$

$$\begin{aligned}
&= \frac{1}{K} \sum_{i=1}^K [A(X_i - m_x)][A(X_i - m_x)]^T = \\
&= \frac{1}{K} \sum_{i=1}^K A(X_i - m_x)(X_i - m_x)^T A^T = \\
&= A \left[\frac{1}{K} \sum_{i=1}^K (X_i - m_x)(X_i - m_x)^T \right] A^T = \\
&= AC_x A^T =
\end{aligned}$$

$$\begin{aligned}
&= \begin{bmatrix} V_1 \longrightarrow \\ V_2 \longrightarrow \\ \vdots \\ V_n \longrightarrow \end{bmatrix} C_x \begin{bmatrix} V_1 & V_2 & \cdots & V_n \\ \downarrow & \downarrow & \downarrow & \downarrow \end{bmatrix} = \\
&= \begin{bmatrix} V_1 \longrightarrow \\ V_2 \longrightarrow \\ \vdots \\ V_n \longrightarrow \end{bmatrix} \begin{bmatrix} \lambda_1 V_1 & \lambda_2 V_2 & \cdots & \lambda_n V_n \\ \downarrow & \downarrow & \downarrow & \downarrow \end{bmatrix} = \\
&= \begin{bmatrix} \lambda_1 & & & \circ \\ & \lambda_2 & & \\ & & \ddots & \\ \circ & & & \lambda_n \end{bmatrix}
\end{aligned}$$

(where the last equality is because the eigen vectors are orthonormal).

2.3.1 Application of the KLT in Geometric Manipulation of Images

One of the applications of the KLT (known also as the Hotelling transform) is in the area of image recognition. Some of the image recognition schemes use the geometry of the image for classification. A problem associated with this method is that the input images may be displaced or rotated relative to the target image they are compared to. The KLT may be helpful in solving this problem for images with a certain geometry. When the object has a geometry in which it seems most elongated along a certain direction (called its 'principal axis'), the KLT may be used to align the image along a known reference axis in a new set of coordinates. The centroid of the image (the average of the x and y coordinates of all pixels) is shifted to the origin and the image is rotated by an angle that minimizes its moment of inertia

[9]. Since the identity of the object to be classified is not known, aligning the image with its principal axis can help decrease errors due to the effects of translation and rotation in the analysis.

2.3.2 Application of the KLT in aligning an image along its principal axis

Next we describe the application of KLT in aligning an image along its principal axis. For a given image, applying the transform in the form that will be presented in this section will result in an image with its centroid at the origin of the new axes and its principal axis (the direction along which it seems most elongated) aligned along the x' axis. This is illustrated in fig. 2.10, where the old coordinates are designated by x, y and the new ones by x', y'

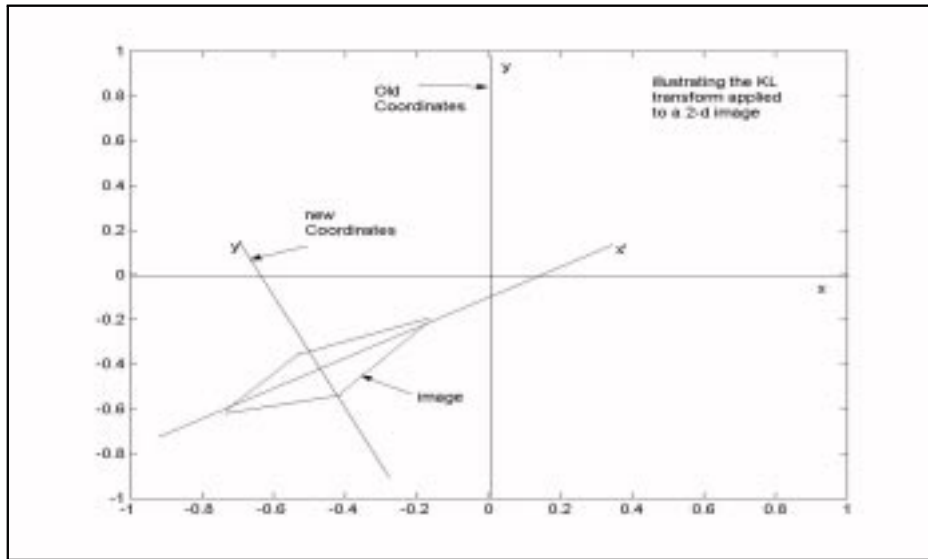


Figure 2.10. Application of the KL to a 2-dimensional image results in a rotated version of the image along its principal axis.

To implement the KLT in this form, we represent each pixel by its x and y coordinates resulting in a population of 2-dimensional vectors (the intensity of the pixels is irrelevant and the image may be considered as a binary image). We compute the covariance matrix (2×2) for this vector population and derive the eigenvalues and the corresponding eigenvectors of the covariance matrix. The pair of eigenvectors (which is an orthogonal set) represents the new set of coordinates. The direction of the

eigenvector corresponding to the larger eigenvalue is equal to the direction of the 'principal axis' of the image.

Given an image of size $k = n \times n$ pixels, let X be a set of k 2-dimensional column vectors X_1, X_2, \dots, X_k , where X_i represents the (x, y) coordinates of pixel i . The mean vector of the population, m_x , is given by:

$$m_x = \frac{1}{K} \sum_{i=1}^K X_i$$

When we compute C_x , the covariance of this vector population is a 2×2 real and symmetric matrix (because the entries in all vectors are pixel coordinates and so they are real values); therefore C_x has 2 real eigenvalues and 2 real orthogonal eigenvectors.

Let λ_1, λ_2 be the eigenvalues of C_x with $\lambda_1 \geq \lambda_2$ and V_1, V_2 be the set of orthonormal eigenvectors, where V_1 corresponds to λ_1 , the largest eigenvalue. Define the matrix A whose rows are the eigen vectors V_1, V_2 :

$$A = \begin{bmatrix} V_1 & - & - & - & - & > \\ V_2 & - & - & - & - & > \end{bmatrix}$$

This matrix represents the new coordinates and can be used as the transformation matrix to align the image along its 'principal axis'. We first move the centroid of the image to the origin of the new coordinates by subtracting the mean of the vector population from each vector and then apply the transformation using the matrix A :

$$Y_i = A(X_i - m_x) \quad i = 1, 2, \dots, k$$

where Y_i represents the (x, y) coordinates of pixel i in the new coordinate system.

2.4 The Approximate Karhunen Loeve Transform

As was mentioned in the previous section, Karhunen Loeve Transform (KLT) provides a global optimum with respect to the entropy cost function for an ensemble of vectors with multivariate normal distribution, but its computational complexity is $O(d^3)$ where d is the dimension of the vectors in the en-

semble. For desktop computation we are limited to $d \leq 10^3$ [33]. We would like to derive a transformation that is computationally feasible and is close to the KLT with respect to an information cost function.

The *approximate Karhunen Loeve Transform* can provide such a basis with computational complexity of $O(d^2 \log d)$, and it will be described following the notation in [33].

Given a vector population $X = \{X_1, X_2, \dots, X_N\}$, a set of N d -dimensional column vectors, we first compute its Joint best basis as was described previously. Let U be the $d \times d$ transformation matrix which corresponds to the computed Joint best basis of X ,

$$U = \begin{bmatrix} U_1 & - & - & - & - & > \\ U_2 & - & - & - & - & > \\ & & & & \cdot & \\ & & & & \cdot & \\ U_d & - & - & - & - & > \end{bmatrix}$$

Assume the rows of $U = \{U_i \in R^d, i = 1, 2, \dots, d\}$ are arranged such that the variance of $(UX) = \{UX_1, UX_2, \dots, UX_N\}$, i.e. the variance of the vector population in the Joint best basis, is in decreasing order. Let $d' < d$ be the smallest integer such that

$$\sum_{i=1}^{d'} \sigma(UX)_i \geq (1 - \epsilon) \text{Var}(X)$$

where $\sigma(UX)_i$ is the variance of the i^{th} coordinate in the Joint best basis, $\text{Var}(X)$ is the total variance of the vector population X , and $\epsilon > 0$ is a constant. Then d' of the first basis vectors in the Joint best basis contain $(1 - \epsilon)$ of the total variance of the vector population X .

Let U' be the $d' \times d$ matrix containing the first d' basis functions of U :

$$U' = \begin{bmatrix} U_1 & - & - & - & - & > \\ U_2 & - & - & - & - & > \\ & & & & \cdot & \\ & & & & \cdot & \\ U_{d'} & - & - & - & - & > \end{bmatrix}$$

The projection (coefficients) of the vector population X onto the d' coordinates of the Joint best basis is given by

$$U'X = \{U'X_1, U'X_2, \dots, U'X_N\}$$

We now apply the Karhunen Loeve transform to these set of coefficients. This can be done by computing the covariance matrix M of the vector coefficients in U' . The mean vector m , in this basis, is given by:

$$m = E[U'X] = \frac{1}{N} \sum_{i=1}^N U'X_i$$

and the covariance matrix, $C_{(U')}$, of the vectors in the basis U' is given by:

$$\begin{aligned} C_{(U')} &= E[(U'X - E(U'X))(U'X - E(U'X))^T] = \\ &= \frac{1}{N} \sum_{i=1}^N (U'X_i)(U'X_i)^T - mm^T \end{aligned}$$

Note that $C_{(U')}$ is a $d' \times d'$ matrix. We compute the eigen vectors of $C_{(U')}$.

Let K be the $d' \times d'$ matrix, where the rows of K are the eigen vectors of $C_{(U')}$. Then

$$K * U'$$

is a $d' \times d$ matrix representing the composition of the Karhunen Loeve Transform applied on the Joint best basis. We define $K * U'$ to be the *approximate Karhunen Loeve Transform*.

Now, as we will see in the experimental results in chapter 4, even for a small ϵ we can obtain $d' \ll d$, decreasing the computational complexity involved in the computation of KLT from $O(d^3)$ to $O(d'^3)$.

2.4.1 Biplot of data using the first two coordinates in a feature vector

For certain vector populations, the first two coordinates of the KLT provide a good representation of the data and, for the purpose of discriminating between classes, if the first two coordinates carry sufficient discriminating information, we can use a simple biplot to visualize the classification results[13]. We will use this idea but choose the approximate Karhunen Loeve Transform for that purpose as it combines the Joint best basis (which provide a good representation of the population) and the KLT, which provides high transform coding gain. We will represent each vector in two-dimensional plot, using the values of the first two coordinates as the x,y coordinates. If the two coordinates can discriminate between the two

classes, we would be able to observe two clusters. We will use a biplot using the first two coefficients in the Joint best basis and the approximate KLT basis.

2.5 Feature Extraction and Classification

The important features in signals of interest such as mammographic images, satellite images or musical recordings are normally localized in space and in frequency. In mammographic images, the characteristics (sharpness, regularity, etc.) of the borders of calcification points and mass lesions are important factors in discriminating between benign and malignant tissues. These characteristics are local both in space and frequency, and the conventional Fourier analysis techniques are not useful in detecting them. Extracting the relevant feature from the class of signals is the first step in classification and it is important especially for the type of signals in our data base. Images of 128x128 pixels cannot be analyzed directly due to their high dimensionality.

2.5.1 Measures of Energy Distributions

Transforming a signal from one base to another using an orthogonal transformation may change the distribution of the signal's energy along the various coordinates, though the energy (using say euclidean norm) does not change. For a population of vectors, the total variance of the ensemble is associated with the energy of the vectors in the ensemble and is invariant under unitary transformation. Unitary transformations are equivalent to the rotation of the coordinate axes in the d dimensional space, and so the total variance is invariant. However, the distribution of the total variance among the d coordinates may change. The 'variance ellipsoid' with semiaxes $\sigma_1^2, \sigma_2^2, \dots, \sigma_d^2$ has a volume proportional to the product of its semiaxes. The volume of this ellipsoid depends on the choice of the coordinate system. KLT minimizes the volume of this ellipsoid for an ensemble of vectors with multivariate normal distribution and is a global minimum among all orthogonal transformations, which implies that KLT provides the highest possible amount of compression for multivariate normal distribution.

We use two measures to compare the distribution of a vector population in different bases, the *transform coding gain* and the *accumulative variance*. They both are based on the variance values, but use different operators.

2.5.1.1 Transform Coding Gain(TCG)

Let $V = \{v_1, v_2, \dots, v_N\}$ be a set of N vectors with d coordinates each. For simplicity, assume the average vector is zero, i.e. $E(V) = \sum_{i=1}^N v_i = 0$. Then the variance of the p^{th} coordinate is given by:

$$\sigma_p^2 = \frac{1}{N} \sum_{i=1}^N v_i^p$$

where v_i^p is the value of the p^{th} coordinate of vector v_i . The *transform coding gain* [26] of the transformation U is given by the ratio of the arithmetic mean of the variances of the coefficients to their geometric mean:

$$G_{TC}(U) = \frac{\frac{1}{d} \sum_{i=1}^d \sigma_i^2}{(\prod_{i=1}^d \sigma_i^2)^{\frac{1}{d}}}$$

Since the sum of variances of the ensemble is invariant under orthogonal transformation, $G_{TC}(U)$ is maximized when the denominator, the geometric mean, is minimized. The geometric mean is proportional to the variance ellipsoid we mentioned above. The KLT provides the highest transform coding gain among all orthogonal transformations; however, in addition to its high computational complexity, (unlike for example, the Fourier transform, or the standard wavelet transform), the KLT is data dependent; the basis functions are derived from the covariance matrix of the data. (This is another reason the KLT is not practical for some applications).

In chapter 3 we will expand the transform coding gain measure to a vector form. We will use the TCG of the KLT basis and the Joint best basis as feature vectors for classification.

2.5.1.2 Accumulation of Variance

The accumulated energy along the coordinates may also be used to compare energy distributions in different bases. For a single vector $V = (c_1, c_2, \dots, c_d)$, the normalized accumulative energy is defined to be:

$$L(k) = \frac{\sum_{i=1}^k c_i^2}{\|V\|^2} \quad k = 1, 2, \dots, d$$

For an ensemble of vectors, let $V = \{v_1, v_2, \dots, v_N\}$ be a set of N vectors with d coordinates each, as before, assuming the average vector is zero, i.e. $E(V) = \sum_{i=1}^N v_i = 0$. Then the variance of the p^{th} coordinate is given by:

$$\sigma_p^2 = \frac{1}{N} \sum_{i=1}^N v_i^p$$

where v_i^p is the value of the p^{th} coordinate of vector v_i . The accumulated variance, $AccVar$, is a vector whose k^{th} entry, $AccVar_{(k)}$, is given by:

$$AccVar_{(k)} = \sum_{p=1}^k \sigma_p^2$$

In general the energy distribution of signals in the original coordinates (i.e. euclidean coordinates) is more balanced than the distribution of the coefficients using a transformation such as the wavelet transform (this feature of the wavelet transform accounts for its compression property). The accumulation of variance will provide a measure of the compression rate of the transformation. It will also be used as a feature vector for classification.

2.5.2 Compression and Approximation Error

Compression of a signal involves expressing the signal using a smaller set of coefficients. This may result in some loss of information. We provide a quantitative measure of the loss when we truncate the coefficients in terms of the total energy of the signal.

We follow the notation in [29]. Let $f(x)$ be a function and $U = \{u_1(x), \dots, u_m(x)\}$, an orthonormal basis. The function $f(x)$ can be represented by the basis functions $u_1(x), \dots, u_m(x)$:

$$f(x) = \sum_{i=1}^m c_i u_i(x)$$

and so in the basis U , $f(x)$ is now represented by the coefficients c_1, c_2, \dots, c_m . Arrange the coefficients in decreasing order

$$|c_{\pi(1)}| \geq |c_{\pi(2)}| \geq \dots \geq |c_{\pi(m)}|$$

where $\pi(i)$ is some permutation of $1, 2, \dots, m$.

We want to represent $f(x)$ with a smaller number of coefficients than the m coefficients. Assume we approximate $f(x)$ by using only $m' \leq m$ largest (in magnitude) coefficients:

$$f_{app}(x) = \sum_{i=1}^{m'} c_{\pi(i)} u_{\pi(i)}(x)$$

The L^2 error is then

$$\varepsilon = \|f(x) - f_{app}(x)\|$$

Now we get an estimate of the square of the error as a function of the number of coordinates we use:

$$\begin{aligned} \varepsilon^2 &= \|f(x) - f_{app}(x)\|^2 = \left\| \sum_{i=1}^m c_i u_i(x) - \sum_{i=1}^{m'} c_{\pi(i)} u_{\pi(i)}(x) \right\|^2 = \\ &= \left\| \sum_{i=m'+1}^m c_{\pi(i)} u_{\pi(i)}(x) \right\|^2 = \sum_{j=m'+1}^m \sum_{i=m'+1}^m c_{\pi(i)} u_{\pi(i)} c_{\pi(j)} u_{\pi(j)} = \end{aligned}$$

and since $u_1(x), \dots, u_m(x)$ is an orthonormal set, we get

$$= \sum_{i=m'+1}^m c_{\pi(i)}^2$$

If we want to approximate $f(x)$ with a L^2 error of not more than ε , we should choose the smallest $m' \leq m$ satisfying

$$\sum_{i=m'+1}^m c_{\pi(i)}^2 \leq \varepsilon^2$$

2.5.3 Fisher's Linear Discriminant Analysis (LDA)

There are many problems that may be computationally manageable for a low dimension, but completely impractical when the dimension goes up to 100. Fisher's Linear Discriminant is one of the tech-

niques used to reduce the dimensionality of a classification problem [3]. It reduces the problem from d dimensions to one dimension. In general given a population of d -dimensional samples, Fisher's Linear Discriminant computes the line (in d -dimensional space) for which the projected samples on this line (which are scalars) are best separated. Fig. 2.11 illustrates Fisher's LDA for 2-d in the case of two classes.

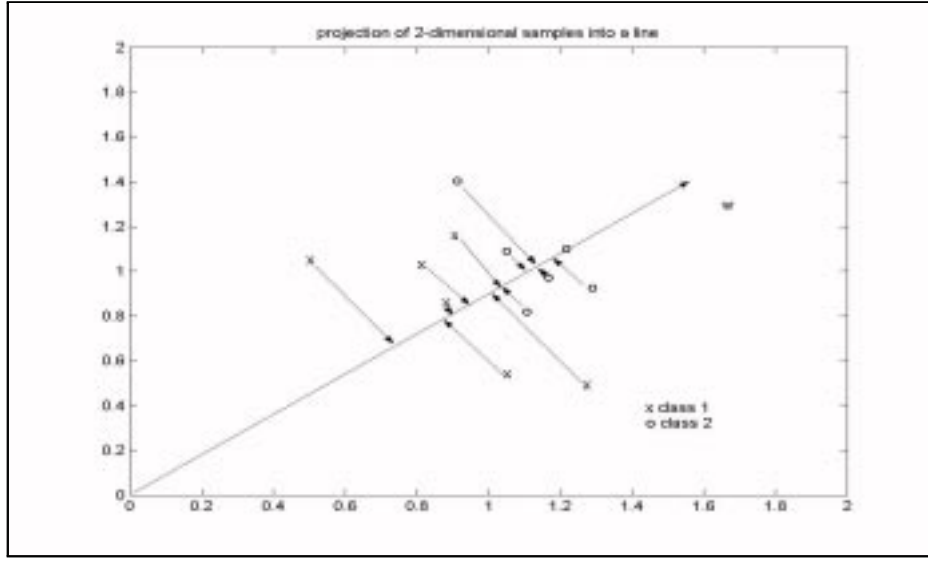


Figure 2.11. Illustration of Fisher Linear Discriminant as a classifier. The projection of the samples onto Fisher's LD are used for classification.

We present Fisher's LDA for the case of 2 classes. Let $X = \{x_1, x_2, \dots, x_n\}$ be a population of n samples (vectors) where $x_i \in R^d$. Assume the samples composed of two classes labeled ω_1 and ω_2 : $X_1 = \{x_1, x_2, \dots, x_{n_1}\}$ and $X_2 = \{x_1, x_2, \dots, x_{n_2}\}$ so that $n = n_1 + n_2$. Let w be a d -dimensional column vector representing a line in the d -dimensional space. The projection of any sample (vector) x_i onto w results in a scalar y_i given by:

$$y_i = w^t x_i$$

and so w^t is a mapping of the sample(vectors) population $\{x_1, x_2, \dots, x_n\}$ onto the scalars $\{y_1, y_2, \dots, y_n\}$.

$$w^t : \{x_1, x_2, \dots, x_n\} \longrightarrow Y = \{y_1, y_2, \dots, y_n\}$$

or in term of the two classes,

$$w^t : \{x_1, x_2, \dots, x_{n_1}\} \longrightarrow Y_1 = \{y_1, y_2, \dots, y_{n_1}\}$$

$$w^t : \{x_1, x_2, \dots, x_{n_2}\} \longrightarrow Y_2 = \{y_1, y_2, \dots, y_{n_2}\}$$

We now present the derivation of the vector (line) w that best provides a separation of the projected samples onto the line. A quantitative measure of how well the two classes are separated is the difference of the sample means. Let m_i be the sample mean of class i :

$$m_i = \frac{1}{n_i} \sum_{x_i \in X^i} x_i \quad i = 1, 2$$

The *scatter matrix* S_i of class i , is defined to be:

$$S_i = \sum_{x_i \in X^i} (x_i - m_i) (x_i - m_i)^t \quad i = 1, 2$$

The *within class scatter matrix* S_w is defined to be:

$$S_w = S_1 + S_2$$

and the *between class scatter matrix* S_B is defined to be:

$$S_B = (m_1 - m_2) (m_1 - m_2)^t$$

Fisher Linear Discriminant maximizes the criterion function $J(w)$:

$$J(w) = \frac{w^t S_B w}{w^t S_w w}$$

It can be shown [3] that the vector w that minimizes $J(w)$ is given by:

$$w = S_w^{-1} (m_1 - m_2)$$

Fisher Linear Discriminant as a Classifier

The mapping $w^t : \{x_1, x_2, \dots, x_n\} \longrightarrow Y = \{y_1, y_2, \dots, y_n\}$ is many to one (i.e. there is more than one population that will have the same projection on the line w). Fisher Linear Discriminant Analysis rarely provides good classification results [3], but it has the advantage that the classification problem is reduced to one dimension. If the distributions of the population in each class are multivariate normal distributions with equal covariance matrices, the Fisher Linear Discriminant does achieve the minimum classification error.

Chapter 3

Methodology

3.1 Combining the Hotelling Transform in Image Query

In this section we describe the method of applying the Karhunen Loeve Transform to align an image along its principal axis and its integration in the wavelet based image query.

3.1.1 Wavelet Based Approach in Image Query

Recently, a group of researchers developed a strategy based on wavelets that provides a fast and efficient image query algorithm in large data bases [12]. In this approach, both the query image and the target image are represented by a small number of quantized wavelet coefficients. Using a metric that is tolerant to large difference between the query image and the target image, the algorithm can perform a comparison that takes into account only the dominant features of the two images. Details of this approach can be found in [12,29].

As wavelet coefficients are not invariant under displacement and rotation of an image, the performance of the search algorithm is very sensitive when the centroids of the two images are not at the same point or if one is rotated relative to the other (for comparison, the scheme that is based on comparing the color histograms of images is not sensitive to displacement and rotation). The KLT can be used to improve the robustness of the wavelet based algorithm in image query. It can be used to lower the error rate due to its sensitivity to displacement and rotation. The improvement however, would involve an increase in the computation time.

3.1.2 Aligning an image along its principal axis

In chapter 2 we described an application of the KLT in the geometric manipulation of images. Given a query image (binary), *we represent each pixel by its x and y coordinates* resulting in a population of 2-dimensional vectors. The covariance matrix, C_x , for this vector population is a (2×2) matrix. We compute its eigen values and eigen vectors.

Let λ_1, λ_2 be the eigenvalues of C_x with $\lambda_1 \geq \lambda_2$ and V_1, V_2 be the set of orthonormal eigenvectors, where V_1 corresponds to λ_1 , the largest eigenvalue. Then the matrix A :

$$A = \begin{bmatrix} V_1 & - & - & - & - & > \\ V_2 & - & - & - & - & > \end{bmatrix}$$

represents the transformation matrix. To displace the image to its centroid and rotate it so that it is aligned along its principal axis, we simply use the following transformation:

$$Y_i = A(X_i - m_x) \quad i = 1, 2, \dots, k$$

where X_i represents the (x, y) coordinates of pixel i in the old coordinate system, m_x represent the average of the coordinates, and Y_i represents the (x, y) coordinates of pixel i in the new coordinate system.

The details of integrating the KLT in the wavelet based image query including experimental results in the context of the alphabet characters (relating to the sensitivity of the wavelet based image query to displacement and rotation and the improvement that can be achieved by applying the KLT) are provided in the published paper in appendix A.

3.2 Enhancement and Classification of Mammographic Images

3.2.1 Mammographic Data Base

We have used a well known data base of mammograms from Nijmegen, the Netherlands, which can be found in the Digital Data Base for Screening Mammography (DDSM) of the University of South

Florida Digital Mammography Home Page. Since 1975 some 13,500 women aged 35-49 years have been invited for breast screening every 2 years in a population based project in Nijmegen, the Netherlands [22]. Information regarding referral, detection and disease stage were calculated and recorded. The mammograms we use come from processed data base and consists of 105 ROI's (regions of interest) contributed by the University of Bologna, Italy, to Dr. Nathan Intrator from Brown University..

Each of the 105 ROI's, is of size 128x128 pixels derived from screen film mammograms with a pixel size of 0.1 mm and a 12-bit gray scale and is large enough to contain a few microcalcifications or the majority of microcalcifications in a cluster. The mammographic images contain 29 benign and 76 malignant . The mammograms are from an early stage and come from general screening of women population and not from x-ray's of women who were sent to take a mammogram due to some pathological indicators (e.g. pain, lump in breast, asymmetry in breasts).

3.2.2 General framework in representing and analyzing an image

Since the number of mammograms is not large, there is no point in analyzing the image as a whole, e.g. with Principal Component Analysis (PCA) or wavelet packet analysis, as the high dimensional space for such representation is extremely sparse. Rather, we would represent each mammographic image by a collection of segments sampled from the image. Most of the background structure in both classes (benign and malignant) is similar, and small segments of size 8x8 pixels are sufficiently large to contain differences relevant to classification (e.g. differences between various characteristics of calcification points such as shape, irregularity, etc.).

Based on these assumptions we will represent each image by a population of 4096 segments each of size 8x8 pixels or segments of size 16x16 pixels, sampled with overlapping regions of 3 pixels on the 4 borders of each segment (except segments along the border of the image).

3.2.3 Indicators for Breast Cancer in Mammographic Images

The two main indicators associated with breast cancer are microcalcification clusters and masses. Microcalcifications appear in mammograms as tiny areas (with a size of a few pixels in digitized images or about .2mm in diameter) that are slightly brighter than the background. Microcalcification clusters are not always easy to detect. They are observed by radiologists in 30%-50% of all malignant mammograms, but in pathological examination, 80% of breast carcinomas contain microcalcifications [2]. Some of the microcalcifications are arterial calcifications which are benign and account for 50% of the false positive detection of malignant calcification [8]. The size and irregularity of a single microcalcification are important features for discrimination. Enhancement of mammographic features can be an important tool to help detect calcification points that are not easily visible to the radiologist. As to a cluster of microcalcifications, the size and circularity of the cluster as well as the number and variation of the size and shape of microcalcifications within the cluster are important features in the diagnosis process. Microcalcifications associated with malignant processes generally have more irregular shapes with fuzzy and spicular boundaries. They are also less uniform in density and size and usually are grouped into multiparticle clusters. Benign microcalcifications on the other hand are usually smoother, more well defined, rounded and uniform in density and size [34]. The differences between malignant and benign microcalcifications are greater and more easily observable when the malignancy is in an advanced stage. It is therefore a difficult challenge to distinguish between the two in the early stage of the breast cancer.

The second indicator of breast cancer are masses. The parenchymal background may make identification of suspected mass lesions difficult. The major performance-limiting factor in visual lesion detection in medical images is largely agreed to be image noise [5]. Features of masses that were found to differentiate benign from malignant masses are [20]: degree of speculation, margin sharpness, density of the mass and texture within the mass. Research using some of these features for classification of masses can be found in [8] and [20]. Most of the false positives reported in [8] and related to masses are due to nodular densities on the film that resemble a mass.

In our analysis, we assume that the various indicators mentioned above will manifest themselves in the wavelet packet coefficients of the segments sampled from an image and that they can be used to discriminate between benign and malignant segments.

3.2.4 Image Enhancement as a Preprocessing Step for classification

Some of the characteristics of microcalcifications and masses differentiating benign and malignant are associated with features such as local contrast, sharpness and smoothness of the contour in a cluster. We would like to enhance these features in the mammographic images. Although the energy of these features is concentrated in the high frequency components of the signal, a global filter is not suitable to enhance those features as there is a large contrast range in the background structure both at the range of various mammographic images and also in different areas of an individual image. We therefore need to apply filtering that should be adaptive to variations between individual images and to variations in background at various portions within the image.

There are various adaptive image enhancement techniques that can be employed to enhance features of interest, to blur or to sharpen edges and remove the background in an image. We experimented with some of these techniques and used in this study one of the variants of *local averaging* image enhancement which will be described later. First we describe briefly some of these image enhancement techniques:

The *median filter* replaces the value of a pixel by the median value of its neighborhood. The effect is the removal of isolated spikes and degradation of fine lines in the image.

The adaptive *Difference of Gaussians (DoG)* filter selectively removes low frequency components. It is the difference of two gaussians, where the central lobe strongly promotes each active location in the image, while the broader negative surround inhibits that location if other strongly activated locations are present around. The effect is an adaptive attenuation of the background structure while emphasizing points of interest in the image. We normalize the Gaussian filter to 0 DC (i.e. no pixel will have a negative value).

Unsharp masking, in general, involves blending of high frequency components with low frequency components to achieve an effect of sharpening, in which case it will enhance local contrast by suppressing the overall brightness range of the image, or blurring effect, depending on the proportion of the components involved. If A represents the source image, and B is the image obtained by applying a local averaging (LPF) to image A , then the unsharp masking can be represented by the formula:

$$C = \gamma \cdot (A - B) + B = \gamma \cdot A + (1 - \gamma) \cdot B$$

If $\gamma \in (0, 1)$ the effect is smoothing the image (a blurring effect). If $\gamma > 1$, the effect is sharpening the details (emphasis on high frequencies). The unsharpening can be implemented by convolving the image with the template t

$$t = \begin{bmatrix} v & v & v \\ v & w & v \\ v & v & v \end{bmatrix}$$

where $v = \frac{1-\gamma}{9}$ and $w = \frac{8\cdot\gamma+1}{9}$

Local averaging involves smoothing the image by reducing local variations in intensities of the pixels. There are few versions of this operation: The pixel value may be replaced by the average value of the pixel values in its neighborhood, or by the average divided by the standard deviation of the pixel values in the neighborhood (in which case the smoothing effect is adaptive to local variations in pixel intensity).

One variant of this technique is conditional local averaging, in which only certain pixels are chosen from the neighborhood, the ones whose value does not differ much from the point. We look for all points y in the neighborhood of the point x such that

$$|I(y) - I(x)| < T \text{ where } T \text{ is a threshold value}$$

where $I(y)$ is the intensity of the neighboring pixel and $I(x)$ is the intensity of the pixel being processed. This operation can be repeated a few times to get an increased smoothing effect. The variant we use in this study, *local image normalization*, consists of 3 steps: removal of the dc component at the image level followed by two local neighborhood normalizations:

We subtract from each pixel the average of all pixels in the 128x128 image. This will remove the dc component at the image level.

We divide the intensity of a pixel by the standard deviation of a neighborhood defined by a clique of 9x9 pixels. This would enhance points where the local standard deviation is low and reduce local variations.

Neighborhood operation: we subtract from each pixel the mean of a neighborhood defined by a clique of 9x9 pixels, then divide the result by the standard deviation of the clique. This step would remove low frequency variations in the image.

In chapter 4 we will present the effects of image enhancement in the spatial domain and the frequency domain, the effect on the distribution of pixels' intensities, the distribution of wavelet packet coefficients and the distribution of the energy of the ensemble of segments sampled from a single image.

Block Processing in MATLAB

The image processing toolbox of MATLAB provides a block processing function which simplifies the implementation of these image enhancement techniques and other operations that have to be applied to multiple overlapping segments in an image. The function *blkproc* takes as an input the image to be processed, the size of the template to be used, the overlapping border and the function to be used on the template. We use this function for image enhancement and other operations which involve processing multiple segments in an image.

In chapter 4 we show the effect of image enhancement in the spatial and frequency domain. Having each image represented by an ensemble of 4096 segments we will also compare the distribution of the energy of the ensemble along the various coordinates of the JBB for the unprocessed and enhanced images.

3.2.5 Feature Extraction

The problem of associating certain features in an image with certain coordinates is not an easy one for high dimensional images. Signal compression can be a useful tool in this context. Good compression

enables a high dimensional signal to be represented by many fewer coordinates that capture most of its energy. In certain cases, those few coordinates may be associated with some features of interest. This can make the task of extracting those features feasible, either for enhancement, alerting a human to those features or using those features in classification. With a smaller number of coordinates it is easier to analyze the relation between these coordinates and features of interest. This is the objective of the two major statistical methods for the purpose of feature extraction: the Fourier transform and the KLT. They both offer an efficient system of coordinates to solve certain problems. The problem with both methods is that when the features of interest have energy localized both in space (or time) and frequency, they fail to capture these features. KLT will capture global features either in the time domain or in the frequency domain but not in both. Additionally, KLT is also sensitive to outliers and has a high computational complexity of $O(n^3)$ where n is the dimension of the problem. The Fourier transform has good localization in the frequency domain but poor localization in the time domain due to the averaging over the full time interval.

As to the relation of the compression property of a basis and its use for discrimination, we note that a basis with the best compression ratio may not be the right basis for discrimination. It is possible that a basis with less compression ratio will have a better correlation of some of its coordinates with the discriminating features in the image. This can be seen in the next figure. Application of the Karhunen Loeve transform to the population (composed of two classes) in the figure will result in the new set of coordinates shown in the figure 3.1. Although in the new coordinates most of the variance of the vector population is along the coordinate x' , it has less discriminating power when compared to the older x coordinate. In the original coordinates, even though the variance of the whole population is approximately evenly distributed between the two coordinates, it is clear that the projection on the x axis can be used to discriminate between the two classes.

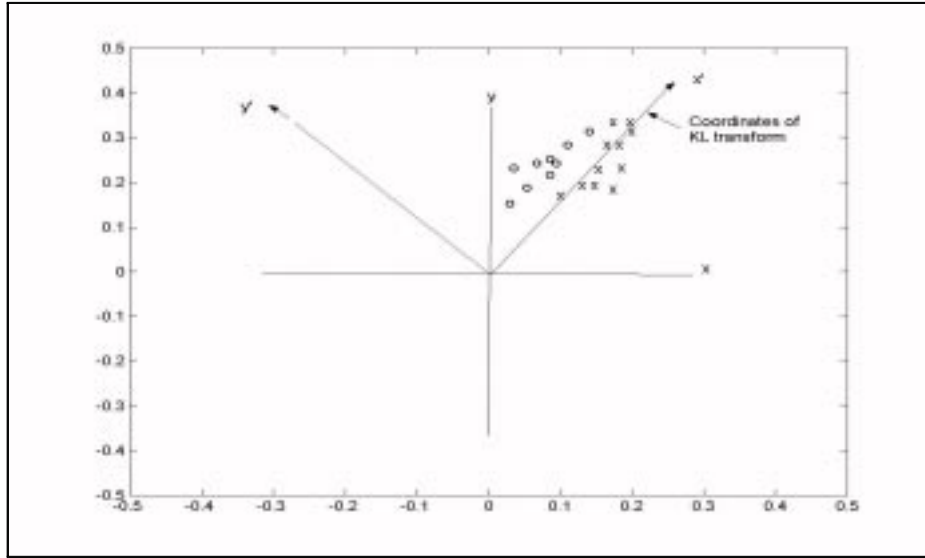


Figure 3.1. The Karhune Loeve coordinates for a population of 2-dimensional vectors consisting of two classes

3.2.6 Justification for Using Feature Vectors Based on Wavelet Packet Coefficients

Wavelets and wavelet packets were found to achieve good compression ratio while preserving fine features in the compressed image due to their good spatial/frequency localization. In chapter 2 we described the best basis algorithm, which for a certain signal searches among a family of basis functions in the wavelet packet table (with respect to a certain wavelet), for a basis which minimizes the entropy of the coefficients. The entropy in this search is appropriate as it provides a measure of the number of dominant coordinates in a vector and the search is efficient and costs $O(n(\log n))$. We also presented the Joint Best Basis, which is the extension of the best basis to an ensemble of signals. It is a basis which minimizes the cost function summed over all the signals in the ensemble. The system of coordinates provided by the Joint Best Basis provides good spatial/frequency localization for the ensemble of vectors it represents. Our features of interest which were described previously are localized both in space and in frequency. We will experiment with a variety of family of bases (using various parameters, e.g. various wavelets, number of segments in an ensemble) and hope that a few will provide a good system of coordinates that will capture the features which discriminate between benign and malignant mammograms. In

addition to the use of the Joint Best Basis to derive feature vectors, we would also use the *approximate KLT*, which will provide both good spatial/frequency localization due to the Joint Best Basis and good energy compaction due to the KLT applied on the wavelet packet coefficients.

Next we describe the various forms of the Joint Best Basis we use in this study.

3.2.7 Application of the Joint Best Basis to Mammographic Images

The Joint best basis as was described in chapter 2, will be applied in this study in three forms:

In the first which we call the *Common Joint Best Basis*, the basis will be derived from a mixed ensemble of benign and malignant segments sampled from 10 benign and 10 malignant mammograms.

In the second form we will derive a pair of Joint best bases, one from an ensemble of segments sampled from 10 benign images and the second basis derived from an ensemble of segments sampled from 10 malignant images.

In the third form, the Joint best basis will be derived for each individual mammogram. In this case the basis is derived from ensemble of segments sampled from a single mammogram.

The motivation behind each of the above applications of the Joint best basis is the following: In the first case, the *common Joint best basis* should capture the dominant features which are common to both classes. This basis will provide a few coordinates that will account for the common background and texture.

In the second case, we assume there are some dominant features within each class but not common to both classes and that each Joint basis will capture the dominant features in each class.

In the third case, we assume there is a great variance between images (whether benign or malignant) so that the Joint best basis derived from a set of images may not be useful for discrimination. In this case we derive a Joint best basis for each individual image and hope that it can be used to provide a 'signature' for that image and that these signatures can be used for classification.

In all three cases, the Joint best bases have a high transform coding gain (higher than the wavelet basis, as will be shown in the next chapter) and therefore can be used for compression and for enhancement of features of interest in the image. The Joint best basis will also be used as the first step of the *approximate KLT* which provides a higher transform coding gain than the Joint best basis and approaches or may even exceed that of the KL basis.

For the purpose of classification, the wavelet packet coefficients extracted from these Joint best bases will be used differently. In the first case we will find the *discriminating power* of each coefficient (the absolute value of the difference of the coefficient's means divided by the sum of variances in both classes). We hope to find a subset of these coefficients with large discriminating power which can be used for classification. The classification of a test image will be done by computing its wavelet packet coefficients using the common Joint best basis (more correctly, the coefficients of a set of segments sampled from the test image), then extracting the subset of the coefficients with the greatest discriminating power and using the distance of this subset to the mean of each class divided by their variances, as the measure for classification.

For the case of two Joint best bases, one for each class, we first compute two sets of wavelet packet coefficients using the Joint best basis for each class, then we compute the distance of each set to the mean of that class. The results of these comparisons can be used to classify the segments sampled from the test image.

3.2.8 Comparison of different bases using their accumulated variance

As was mentioned in chapter 2, an orthogonal transformation of a signal from one base to another, changes the distribution of the signals' energy along the various coordinates while the sum of the variances is unchanged. The accumulated energy along the coordinates can be used to compare the change in distributions in different bases.

For a single vector $V = (v_1, v_2, \dots, v_d)$, the normalized accumulative energy is defined to be:

$$L(k) = \frac{\sum_{i=1}^k v_i^2}{\|V\|^2} \quad k = 1, 2, \dots, d$$

In general the energy distribution of signals in the original coordinates (i.e., euclidean coordinates) is more balanced than the distribution of the coefficients using a transformation such as the wavelet transform (this feature of the wavelet transform accounts for its compression property).

For an ensemble of vectors, we consider the variance of *each coordinate* and use the accumulated variance to compare different bases.

We represent each mammographic image by an ensemble of 4096 segments of size 8x8 pixels as was described previously. We will derive the accumulated variance of the following bases:

The accumulated variance in the original basis which is the accumulated variance of pixels' intensities.

The accumulated variance in the wavelet basis. This is the accumulated variance of the wavelet coefficients of the ensemble (using the standard wavelet basis which is independent of the data).

The accumulated variance in the Joint best basis. Having an ensemble of segments (either all benign, malignant or mixed), we compute its Joint best basis (as was described in chapter 2). We use that basis (which is a function of the data) to derive a set of coefficients from the standard deviation table of the wavelet packet coefficients. The square of this set of coefficients represent the variance of the ensemble along the coordinates of the Joint best basis.

The accumulative variance in the KL basis. This is given by summing the eigen-values of the covariance matrix of the segments, as was presented in chapter 2. The KL basis is also data dependent (the eigen vectors comprising the transformation matrix are derived from the covariance of the ensemble of vectors).

The accumulated variance in the approximate KL basis. If KL is applied to the wavelet packet coefficients of the ensemble, using its Joint best basis, we would get the accumulation of variance in the approximate KL basis.

3.2.9 Classification Framework for Mammographic Images

The classification framework involves two steps. The first is to find a transformation that provides coefficients that can be used to discriminate between benign and malignant images. We call this set of coefficients a *feature vector*. The best coordinate system (for discrimination), will present the vector population in the *feature space*, as a set of classes which are maximally separated point clouds in R^n . The class separability index is maximized. There are various classifiers that can be used. We use the following:

Fisher's Linear Discriminant Analysis (LDA), which reduces the classification problem from d to 1 (where d in this case is the dimension of the feature vector). LDA is optimal for the case of two classes, when each class satisfies a multivariate normal distribution. But otherwise its performance may be very poor.

K-nn, K-Nearest Neighborhood Classifier, the multivariate version.

These classifiers will be described later in detail in sections 3.2.11.2 and 3.2.11.3 .

3.2.10 Feature Vectors

We use a few feature vectors in this study. Since we represent each image by an ensemble of 4096 segments, we use the energy distribution of the ensemble in different bases along the various coordinates as feature vectors. The wavelet packet atoms of a segment represent information which is localized in space and frequency and combined with the best basis (or in the case of a group of segments, the Joint best basis) algorithm, we may achieve a basis that captures some features which can be used for classification. We describe each of the feature vectors we use:

Computing wavelet packet coefficients for a single class. We compute a common Joint best basis using ensemble of 4096x20 segments, sampled from 10 images from each class; then for each image we derive a feature vector in the following way: we use the Joint best basis to compute the wavelet packet coefficient for each segment and then compute the variance of all 64 coefficients. This provides an 'average' measure of the energy distribution of the wavelet packet coefficients of all segments with respect to the common Joint best ba-

sis. These coefficients can be used as feature vectors for the K-nn classifier. If there are some discriminating features between the benign and malignant segments, they may be represented by some dominant coefficients in the feature vectors and can be used for discrimination.

Computing wavelet packet coefficients for two classes. In this case we derive a Joint best basis for each class (using 10 images for each class). We use the two Joint best bases to compute the wavelet packet coefficient for each segment in each base, then compute the variance of all 64 coefficients for each set. This provides an 'average' measure of the energy distribution of the wavelet packet coefficients of all segments with respect to each of the Joint best bases. We then use these two sets as a feature vector, and apply the K-nn algorithm for classification.

Computing wavelet packet coefficients for a single image. Here we derive a Joint best basis for each image separately, then use the coefficients derived from the standard deviation table of the wavelet packet coefficients as feature vectors.

Using a Distance Measure. Rather than using the two sets of coefficients as was described before, we can use the distance of each set of coefficients to the average of each class (resulting in two distance values) and use this values for classification.

Transform Coding Gain(TCG) as a Feature Vector. As was described in chapter 2, one measure of the energy compaction achieved by an orthonormal transformation is the Transform Coding Gain(TCG) [26]. This measure can be used as a feature vector for the classification of the mammographic images. We hope that if there is some slight but consistent (within a class) difference in the distribution of the variance along the various coordinates, this difference will be captured by the accumulated variance vector or the transform coding gain. Having an image represented by an ensemble of 4096 segments, we will compute the transform coding gain of two bases: the KLT basis and the Joint best basis for this ensemble..

Formally, if $\{\sigma_1^2, \sigma_2^2, \dots, \sigma_d^2\}$ are the normalized ($\sum_{i=1}^d \sigma_i^2 = 1$) variance values in the KLT basis, the n^{th} partial transform coding gain for the KLT basis, $G_{TC(KLT)}^n$, is defined to be:

$$G_{TC(KLT)}^n = \frac{\frac{1}{n} \sum_{i=1}^n \sigma_i^2}{(\prod_{i=1}^n \sigma_i^2)^{\frac{1}{n}}} \quad n = 1, 2, \dots, d$$

Now define the *transform coding gain vector* for the KLT basis, $V_{TC(KLT)}$ to be:

$$V_{TC(KLT)} = \{G_{TC(KLT)}^1, G_{TC(KLT)}^2, \dots, G_{TC(KLT)}^d\}$$

and similarly, if $\{\sigma_1^2, \sigma_2^2, \dots, \sigma_d^2\}$ are the normalized ($\sum_{i=1}^d \sigma_i^2 = 1$) variance values in the Joint best basis, we define the n^{th} *partial transform coding gain* for the Joint best basis,

$G_{TC(jointbb)}^n$ to be:

$$G_{TC(jointbb)}^n = \frac{\frac{1}{n} \sum_{i=1}^n \sigma_i^2}{(\prod_{i=1}^n \sigma_i^2)^{\frac{1}{n}}} \quad n = 1, 2, \dots, d$$

and $V_{TC(jointbb)}$, the *transform coding gain vector* for the Joint best basis of the segment population is defined as:

$$V_{TC(jointbb)} = \{G_{TC(jointbb)}^1, G_{TC(jointbb)}^2, \dots, G_{TC(jointbb)}^d\}$$

Each of these vectors can be used as a feature vector for a mammographic image.

Accumulated Variance in the KLT basis and the Joint best basis as Feature Vectors. Similar to the definition of the transform coding gain, we define the *accumulated variance in the Joint best basis* for a mammographic image:

$$V_{AccVar(jointbb)} = \left\{ \sum_{i=1}^1 \sigma_i^2, \sum_{i=1}^2 \sigma_i^2, \dots, \sum_{i=1}^d \sigma_i^2 \right\}$$

where σ_i^2 is the variance of the i^{th} coordinate in the Joint best basis and the *accumulated variance in the KLT basis*:

$$V_{AccVar(KLT)} = \left\{ \sum_{i=1}^1 \sigma_i^2, \sum_{i=1}^2 \sigma_i^2, \dots, \sum_{i=1}^d \sigma_i^2 \right\}$$

where σ_i^2 is the variance of the i^{th} coordinate in the KLT basis.

We will use these two vectors as feature vectors for the classification of the mammographic images.

The accumulated variance and the transform coding gain are related to each other as they are both functions of the variance values. The rationale behind using both is that by applying different mathemat-

ical operations to the variance values, we may capture a unique 'signature' in the distributions that can be used to discriminate between benign and malignant segments. We hope that if there is some slight but consistent (within a class) difference in the distributions of the coefficients, it may be captured either in the partial sum of the variances, or in the partial product of the variances.

3.2.10.1 Discriminating Power of Coefficients in a Feature Vector

For high dimensional data it may be too expensive computationally to use all the coefficients in the feature vector. Also, since many of the coefficients may not be relevant to the task of discrimination due to very low value of *discriminating power*, we therefore need to extract those coordinates with the largest *discriminating power*.

Given a set of feature vectors for classes 1 and 2, and assuming that both have the same number of coefficients, we define the vector \bar{D} , where its i^{th} coordinate, \bar{D}_i , is the *discriminating power* of the i^{th} coefficient:

$$\bar{D}_i = \frac{|mean(c_i^{class1}) - mean(c_i^{class2})|}{var(c_i^{class1}) + var(c_i^{class2})}$$

$mean(c_i^{class1}), var(c_i^{class1})$ are the mean and variance respectively of the i^{th} coefficient in class 1 and $mean(c_i^{class2}), var(c_i^{class2})$ are the mean and variance respectively of the i^{th} coefficient in class 2.

For high dimensional problems we can use only a subset of the coefficients with the largest discriminating power. The results provided in this study are for segment size of 8x8 pixels which constitute a feature vector of only 64 coefficients. As this dimension is low, we will use all the coefficients in the feature vectors.

3.2.11 Classifiers

The two classifiers we use in this study are the k-nn classifier and Fisher's Linear Discriminant. Having each image represented by a feature vector, we use these classifiers to classify the mammograms as benign or malignant.

3.2.11.1 K-Nearest Neighborhood (k-nn) Classifier

We use the multivariate K-Nearest-Neighborhood (k-nn) classifier. The k-nn algorithm first places all the training points in the feature space. To classify a test point, it examines the local neighborhood of the test point in the feature space to determine which training points are the closest to the test point. It then conducts a class vote among the nearest k neighbors to the test point and assigns the result of this vote to the test point (See Duda&Hart). To achieve a robust estimate of the classification results, we use the Jackknife method. We will run a large number of experiments in which 70% of the data is chosen randomly to serve as training data and the rest as test data. This would provide a better performance estimate of the classifier and the feature vector used. The results of the K-nn algorithm we use are summarized in three parameters:

1. The *average error of misclassification*, defined to be the average percentage of benign misclassified as malignant and malignant misclassified as benign.
2. The *sensitivity* of the classification, defined to be the percentage of malignant classified correctly.
3. The *specificity* of the classification, defined to be the percentage of benign classified correctly.

3.2.11.2 Classification of Mammograms Using the K-nn Algorithm

The K-nn algorithm we use in this study accepts as an input a matrix where each row represents a feature vector corresponding to an image and the class labels for each image (0-benign,1-malignant). It randomly selects 70% of the images as training data and the rest are used as test data. As feature vectors we use either vectors with 64 coefficients or 128 coefficients. The K-nn algorithm will sort the coefficients according to their power of discrimination and uses a predefined subset of the coefficients. To test how the classification performance depends on the number of coefficients (i.e. if there is an improvement in the classification results when we use more coefficients), we ran a few experiments using 10,20,30 of the 64 coefficients with the largest discriminating power.

Figure 3.2 illustrates the graphical results of the K-nn algorithm that was used with 10,20,30 coefficients. In this example there is no improvement when taking more than the first 10 coefficients; therefore one can limit the experiments to the first 10 coefficients.

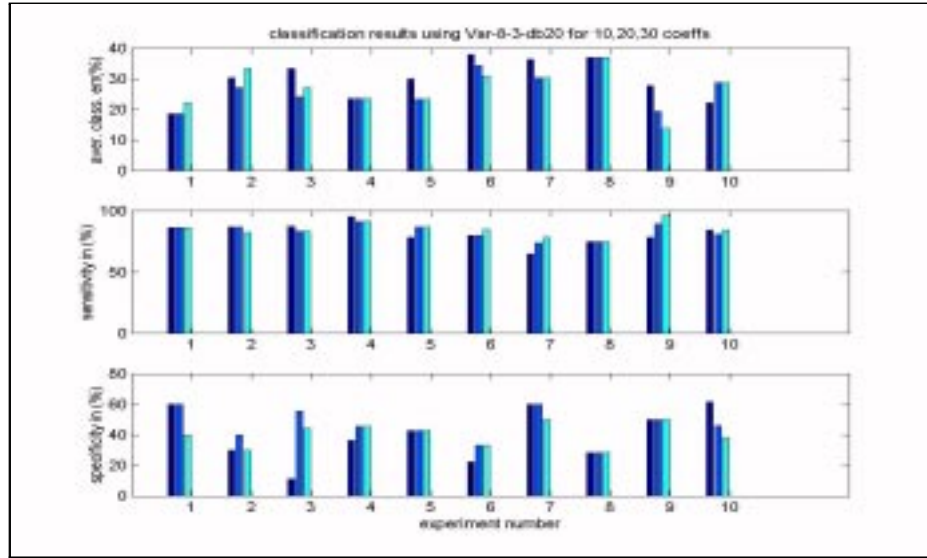


Figure 3.2. Classification results using the Knn classifier for 10 experiments, each with 10,20,30 coefficients: top panel-average error, mid panel-sensitivity, bottom - specificity.

To evaluate the performance of the classification for a certain feature vector, we will run 50 experiments and average (over all experiments) the *average error of misclassification*, the *sensitivity* of the classification, and the *specificity* of the classification. To compare the performance of the various feature vectors, we compare their average results.

3.2.11.3 Fisher's Linear Discriminant Analysis(LDA)

In chapter 2 we described *Fisher's Linear Discriminant Analysis* used to reduce a d -dimensional classification problem to one dimensional classification problem. It requires computing the vector w , *Fisher's linear discriminant*, given by

$$w = S_w^{-1}(m_1 - m_2)$$

where:

m_i is the sample mean of the feature vector of class i :

$$m_i = \frac{1}{n_i} \sum_{x_i \in X_i} x_i \quad i = 1, 2$$

S_i , is the *scatter matrix* of feature vectors of class i :

$$S_i = \sum_{x_i \in X_i} (x_i - m_i) (x_i - m_i)^t \quad i = 1, 2$$

and S_w is the *within class scatter matrix*:

$$S_w = S_1 + S_2$$

Fisher linear discriminant, w , is the linear function that maximizes the ratio of the between-class scatter to within-class scatter.

We compute *Fisher linear discriminant*, w , using the feature vectors of the benign and malignant segments. Then for each feature vector x_i , the *scalar* y_i , is the projection of feature vector x_i onto the line w :

$$y_i = w^t x_i$$

We will apply *Fisher's Linear Discriminant*, w , to all feature vectors of both benign and malignant segments then check the distribution of $\{y_i\}$ for both classes to see how well they are separated. If the distributions are well separated, (checking the distributions visually or using a criterion such as the absolute value of the difference of the means divided by the sum of their variance) then the projection $\{y_i\}$ can be used for discrimination. As was mentioned, LDA has an optimal performance for the case of two classes when both classes satisfy multivariate normal distribution.

3.2.12 Classification Using the *Joint Best Basis*

The Joint best basis will be used in different forms for classification as was described previously. The details will now be presented.

3.2.12.1 Classification using the *common Joint best basis*

We derive the *common Joint best basis* as was described previously using 10 benign and 10 malignant images to create 20x4096 segments of size 8x8. We extract the wavelet packet coefficients of all segments using this single basis.

Let $\bar{\mu}_1, \bar{\sigma}_1$ be the mean and variance vectors of the wavelet packet coefficients extracted from an ensemble of benign segments and $\bar{\mu}_2, \bar{\sigma}_2$ be the mean and variance vectors of the wavelet packet coefficients extracted from an ensemble of malignant segments. Then, the vector \bar{D} , represents the discriminant power of all coefficients and is given by:

$$\bar{D} = |(\bar{\mu}_1 \text{ .- } \bar{\mu}_2) \text{ . / } (\bar{\sigma}_1 \text{ .+ } \bar{\sigma}_2)|$$

where .- , .+ , ./ represent subtraction, addition and division of vectors, element by element and $||$ is the absolute value operator. The i^{th} coordinate of \bar{D} is the discriminating power of coefficient i . We sort the coordinates in \bar{D} in decreasing order so that the first coordinate in the sorted vector represents the coordinate with the greatest discriminating power. For a high dimensional problem we can use only a subset of the coefficients with the largest discriminating power. In this study we work with small segment size (8x8 pixels), resulting in only 64 coefficients and so we will use all the coefficients.

To classify a test image, we first construct an ensemble of 4096 segments sampled from the image and for each segment we apply the following steps: we compute its wavelet packet coefficient using the common Joint best basis, then we find the distance (euclidean norm) from the mean value of each class divided by its variance. If the coefficients are in vector \bar{V} , we compute d_i , its distance to class i , using the following:

$$d_i = \left\| (\bar{V} \text{ .- } \bar{\mu}_i) \text{ . / } \bar{\sigma}_i \right\| \quad i = 1, 2$$

where $|||$ is the euclidean norm. We then decide to classify the segment to be benign if $d_1 < d_2$. Having the classification results, we can classify the image based on the majority of the classification results of its segments (there are other, alternative variants for measuring the distance d_i and its use for classification of the image)

3.2.12.2 Classification using two *Joint best bases*

In this form we compute a Joint best basis for each class using 10x4096 segments to compute each basis. We then extract the wavelet packet coefficients of all segments for each of the classes using its own basis and proceed similar to what we did with the common Joint best basis.

Let $\bar{\mu}_1, \bar{\sigma}_1$ be the mean and variance vectors for the vector coefficients of the benign segments (using the Joint best basis of the benign ensemble) and $\bar{\mu}_2, \bar{\sigma}_2$ be the mean and variance vectors for the vector coefficients of the malignant segments.

To classify a test image, we construct an ensemble of 4096 segments sampled from the image and for each segment we compute its wavelet packet coefficient using the two bases.

Let \bar{V}_1, \bar{V}_2 be the wavelet packet coefficient vectors of a segment derived using the pair of Joint best bases. We find the distance (euclidean norm) from the mean value of each class divided by its variance:

$$d_i = \left\| (\bar{V}_i - \bar{\mu}_i) / \bar{\sigma}_i \right\| \quad i = 1, 2$$

and we decide to classify the segment to be benign if $d_1 < d_2$. The classification of the image can be done as above using the classification results of the majority of the segments.

3.2.12.3 Classification using individual feature vectors

We will also experiment with feature vectors derived for each mammogram. We use the accumulative variance in the Joint best basis, the KLT basis and the approximate KLT basis as feature vectors. These feature vectors are derived individually for each mammogram and they will be used with the k-nn classifier.

3.2.13 Application of the Approximate Karhunen Loeve Transform to Mammographic Images

We apply a variant of the Approximate Karhunen Loeve transform developed by Wickerhauser [33], which was described in chapter 2. Given an image, we will apply the approximate Karhunen Loeve

transform to a population of 4096 segments of size 8x8 sampled from the image. The accumulated variance in this basis will be used as a feature vector for classification

We would also use the first two coefficients in this basis for a biplot of the images as this transform is expected to have a high transform coding gain (close to that of the KLT) to test whether these two coordinates may alone be useful for discrimination.

Let $X^{wp} = \{X_1^{wp}, X_2^{wp}, \dots, X_{4096}^{wp}\}$ be the wavelet packet coefficient vectors of the 4096 segments sampled from an image derived using its Joint best basis.

Let \bar{m} be the expected value of the wavelet packet coefficient vectors of the segment population

$$\bar{m} = \frac{1}{4096} \sum_{i=1}^{4096} (X_i^{wp})$$

Then the covariance matrix of the wavelet packet coefficient vectors is given by:

$$\begin{aligned} M &= E[(X^{wp} - E(X^{wp}))(X^{wp} - E(X^{wp}))^T] = \\ &= \frac{1}{4096} \sum_{i=1}^{4096} (X_i^{wp})(X_i^{wp})^T - \bar{m}\bar{m}^T \end{aligned}$$

For a segment size of 8x8, the covariance matrix is of size 64x64 and it has 64 eigen values. These eigen values represent the variance in the approximate KL basis composed of the Joint best basis followed by KLT.

We use the accumulated variance (which is just the partial sum of the eigen values) to represent each image and apply the K-nn algorithm for classification.

3.2.13.1 Plotting data with respect to the first two coordinates of the approximate KL basis

The main purpose of the KLT is reducing the dimensionality of the problem. When most of the energy in the signal is concentrated in two of the principal components, the data can be represented graphically using its first two principal components (for details see I.T.Jolliffe, *Principal Component Analysis*). We adapt this approach but use the first two coefficients in the approximate KL basis. This basis may prove to be better for classification as it may carry a 'signature' of the image due to the fact that it

is based on the Joint best basis of the segments sampled from the image and it has a high transform coding gain due to the application of the KLT to the wavelet packet coefficients. As we show, the approximate KLT in this study has a transform coding gain almost equal to that of the KLT (which is applied directly on the segment population) and it is far better than the wavelet basis. We would hope that the combination of high transform coding gain and a basis adapted to the segment population of the image will provide a unique signature for the image which can be used for classification.

Chapter 4

Experimental Results

In this chapter we present the results of the classification experiments done with the mammographic images as was described in chapter 3. We first demonstrate the importance of image normalization for the data base we use, then compare the accumulated variance in various bases as a measure of their energy compaction. We present the classification results using the various feature vectors described in chapter 3 and the two classifiers we use, k-nn and Fisher's LDA, and conclude with a summary of the results in which we present the best performance we achieved in this study.

4.1 Effects of image enhancement

Image enhancement is the first step we apply to improve the performance of classification , as was described in chapter 3. We present the effects of the local image normalization (described in section 3.2.4) applied to our mammograms in the spatial domain, the frequency domain, in the distribution of pixels' intensities and in the distribution of the energy of the wavelet packet coefficients of ensemble of segments sampled from the image. In figure 4.1, we present the first malignant mammogram, both the unprocessed and the enhanced version along their frequency spectrum. First note that the strong white background structure is attenuated in the enhanced version emphasizing the details. In the frequency domain, the very low frequency components in the unprocessed image are dominant. The low frequency components account for the background structure. The high frequency components account for the details including the characteristics of the calcification points (e.g. intensity, sharpness, smoothness or irregularity in the border line of the calcification points) and border lines of masses. The enhancement distributes the energy of the image more evenly along the frequency spectrum. We hope that the discriminating features (between benign and malignant) translate to some consistent (within a class) difference in some of the

wavelet packet coefficients associated with the high frequency components and therefore we would like to increase (adaptively) their weight.

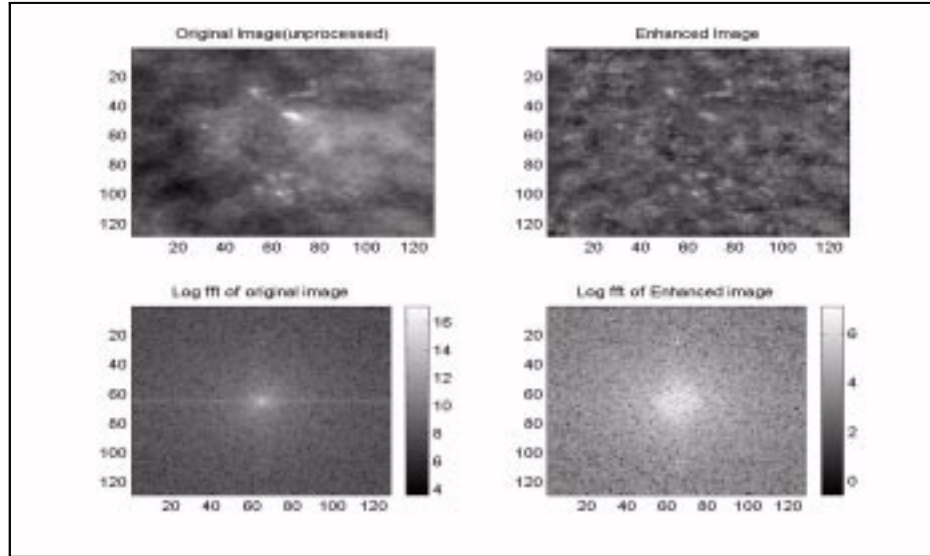


Figure 4.1. Unprocessed and enhanced images and their frequency spectrum

The effect of image enhancement can also be seen in the histogram of pixels' intensities. Fig. 4.2 shows the histogram plots of the third benign and malignant mammograms, both for the unprocessed and the enhanced images. The enhancement has a normalization effect on the distribution of the pixel intensities of the image. Note that the distributions of the enhanced versions are very close to normal with zero mean.

Image enhancement affects also the distribution of the wavelet packet coefficients of the ensemble of segments sampled from an image. Figure 4.3 shows the accumulated variance of the wavelet packet coefficients derived from the common standard deviation packet table using the Joint best basis computed for the first benign mammogram both for the enhanced and the unprocessed image, using 4096 segments each of size 8x8 pixels. In the unprocessed image, most of the energy of the segments will be contained in the few wavelet packet coefficients associated with the low frequency components of the signal. The enhancement has an effect of distributing the energy more evenly over a wider range of coefficients in the

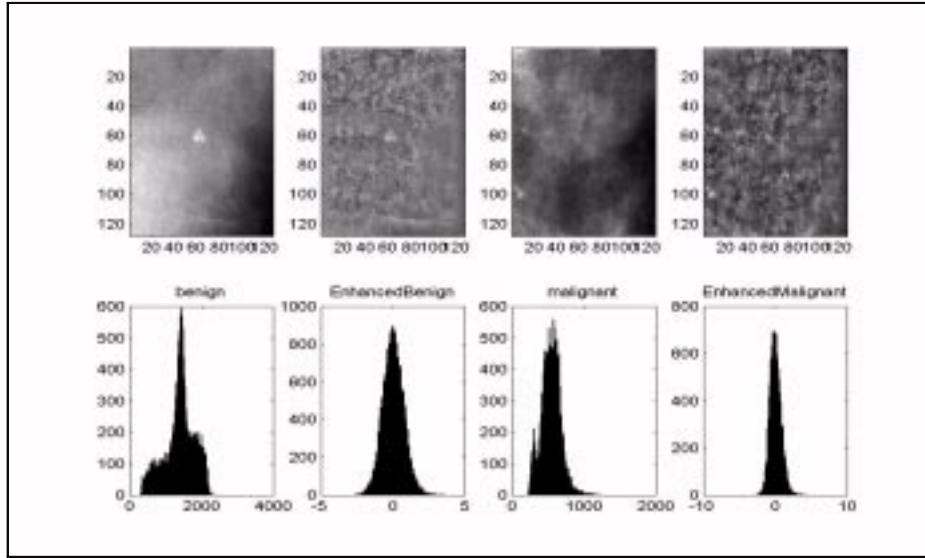


Figure 4.2. histogram of pixel densities for enhanced and unprocessed images

common packet table. This in turn is reflected in a smoother graph representing the accumulated variance in the Joint best basis of the image. This is consistent with the effects in the frequency domain and in the distribution of pixels' intensities. The spreading of the energy will play an important role in our ability to discriminate between benign and malignant mammogram as will be seen later in the classification results.

4.2 Comparing the Accumulated Variance in Various Bases

The accumulated variance provides a measure of the compaction property of a basis. We use it to compare various bases and specifically would like to compare the approximate KLT (Karhunen Loeve Transform) basis to the KLT basis and the Joint best basis. As was described in chapter 2, the approximate KLT is a composition of the KLT and the Joint best basis. We use the KLT to decorrelate the wavelet packet coefficients of the ensemble of segments sampled from an image using its Joint best basis. The KLT is optimal among all orthogonal transformations in terms of the transform coding gain for a population with a multivariate normal distribution. The Joint best basis is already a good basis for the ensemble of segments (better than the standard wavelet transform). The approximate KLT will further decorrelate the

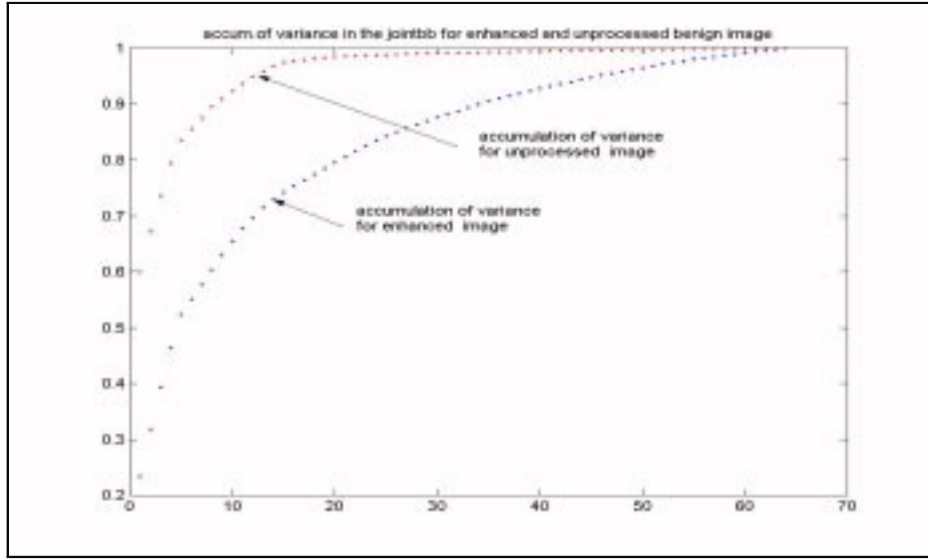


Figure 4.3. accumulation of variance in the Joint best basis for enhanced and unprocessed image. x axis: coefficient number.(64 coefficients for segments of 8x8 pixels)

wavelet packet coefficients in the Joint best basis. The transform coding gain of the approximate KLT should be very close to that of the KLT (which is applied directly on pixels' intensities) and in certain cases (when the distribution is not multivariate normal distribution) it may even exceed that of the KLT.

To evaluate the transform coding gain of the approximate KLT in the context of our mammographic images, we compare the accumulated variance in 5 different bases for 4 mammograms. We present the results for the first benign and first malignant mammograms in the data base, both for the unprocessed and the enhanced versions.

Figures 4.4 , 4.5, 4.6, 4.7 provide the graphs representing the results. Each graph is based on 4096 segments of size 8x8 and the variance values are normalized. The results are given for the original basis, the wavelet basis, the Joint best basis, the approximate KLT and the KLT basis. The wavelet basis is the standard wavelet basis and it is independent of the data while the other bases are data dependent.

First, notice that the accumulation of variance in the original basis is close to a straight line for both the unprocessed and the enhanced images, implying the variance is distributed evenly along all coordinates. Also for all mammograms, the approximate KLT and the KLT bases have about the same

transform coding gain. Both have a better transform coding gain than the Joint best basis and the wavelet basis and the Joint best basis has a better transform coding gain than that of the wavelet basis. In some of the experiments we conducted with larger segments, there were certain cases in which the approximate KLT had a slightly better transform coding gain than that of the KLT.

These results are compatible with the theory presented in chapter 2. In this experiment the approximate KLT achieves a transform coding gain almost equal to that of the KLT and is far better than the standard wavelet transform. For example, for the unprocessed benign mammogram, the first 10 coefficients (out of 64) account for 85% of the variance in the wavelet basis, 90% in the Joint best basis and 95% in the KLT basis and the approximate KLT basis. For the enhanced version of the same mammogram (where the distribution is more even due to image enhancement), the first 10 coefficients account for slightly over 60% of the variance in the wavelet basis, 65% of the variance in the Joint best basis and 70% of the variance in the KLT and the approximate KLT bases.

It is important to note that while the KLT may be the best basis for representation (in terms of energy compaction), it may not be the best tool for classification.

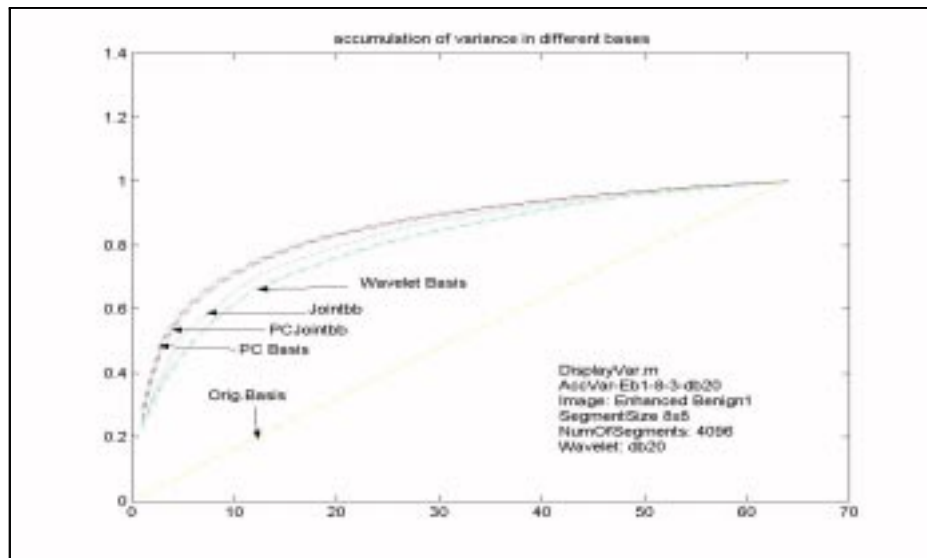


Figure 4.4. Accumulation of variance in different bases for the first enhanced benign mammogram. x axis: coefficient number(64 coefficients for segments of 8x8 pixels)

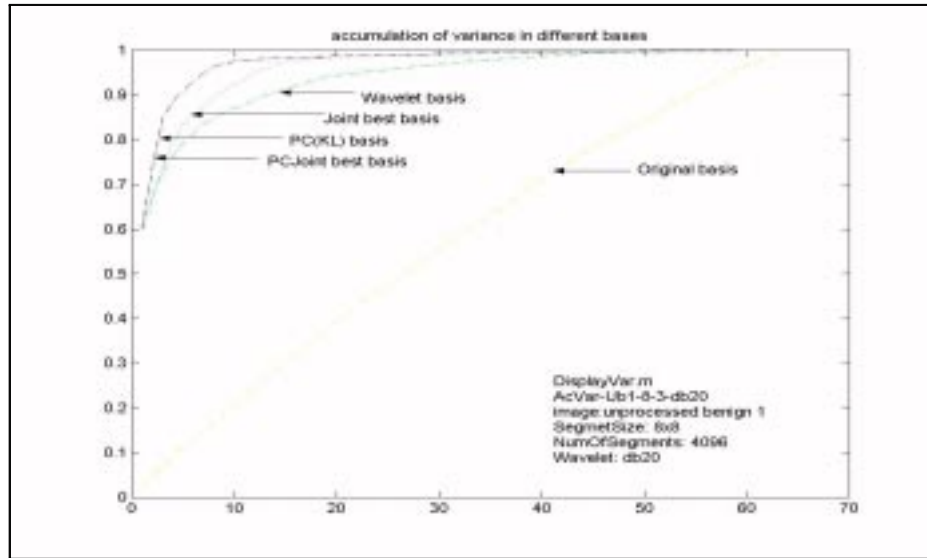


Figure 4.5. Accumulation of variance in different bases for the first unprocessed benign mammogram. x axis: coefficient number (64 coefficients for segments of 8x8 pixels)

The accumulated variance is found in this study to provide the best feature vector for discrimination. As we present later, the accumulated variance provide a good 'signature' for discriminating benign from malignant mammograms.

We also experimented with larger segment size to find out how the KLT and Joint best basis relate to each other for benign and malignant mammograms. For a segment size of 16x16 pixels we found that for most malignant mammograms, the curve representing the accumulated variance in KLT basis was below the curve representing the accumulated variance in the Joint best basis, while for most benign images the relation between the two curves was reversed. The multivariate distribution of benign images is 'closer' to normal than that of a malignant image. The malignancy in the mammogram introduces structure in the image, degrading the normality of some of the coordinates. Since the KLT basis provides the basis with the highest transform coding gain among all orthogonal bases for multivariate normal distributions, the graph representing the accumulated variance in the KLT basis for a benign image will have a higher transform coding gain than that of the accumulated variance in the Joint best basis. This is shown in fig 4.8. The Joint best basis is a 'good' basis when there is some structure in the vector population and so in

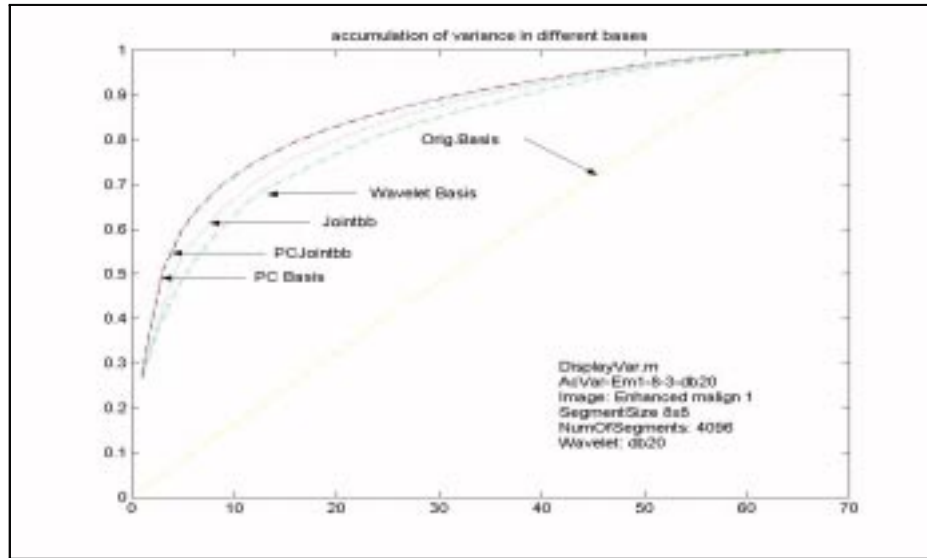


Figure 4.6. Accumulation of variance in different bases for the first enhanced malignant mammogram. x axis: coefficient number (64 coefficients for segments of size 8x8 pixels)

this respect malignant images produce 'better' (in terms of transform coding gain) Joint best basis than the benign mammograms. This result by itself was found to be insufficient to provide robust discrimination between the two classes.

4.3 Classification results

In this section we provide the classification results for the various feature vectors used in this study combined with Fisher's LDA and the k-nn classifiers. We present the results starting with the worst and ending with the best results.

4.3.1 Classification Results Using Fisher's Linear Discriminant

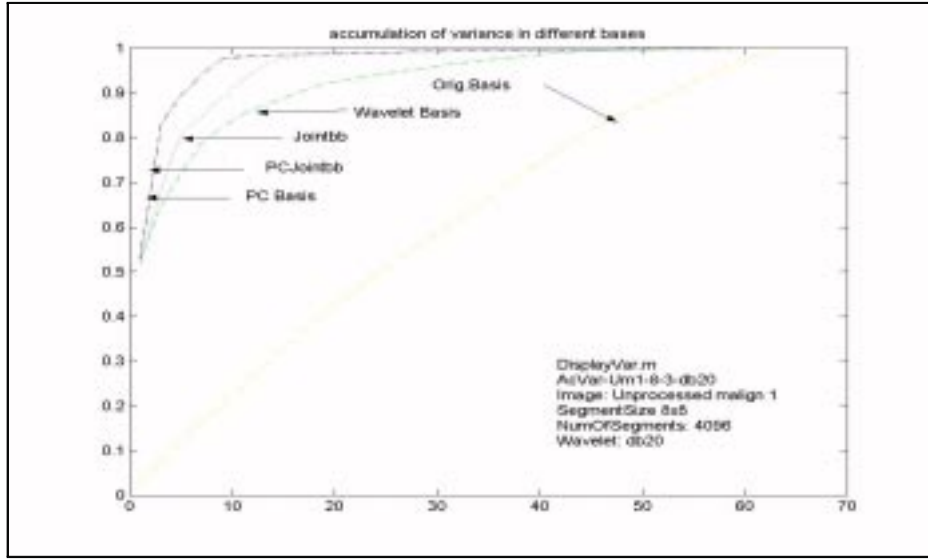


Figure 4.7. Accumulation of variance in different bases for the first unprocessed malignant mammogram (x axis: coefficient number. 64 coefficients for segments of size 8x8)

4.3.1.1 Feature Vectors Based on the Common Joint Best Basis of Enhanced Mammograms

In the next experiment we apply Fisher's Linear Discriminant to the wavelet packet coefficients using the common Joint best basis. We calculated *Fisher's Linear Discriminant*, w , as was described in chapter 3, using as feature vectors the wavelet packet coefficients (extracted from a common Joint best basis) of a large collection of mixed segments sampled from 10 benign and 10 malignant images. We mapped each feature vector x_i into a *scalar* y_i given by:

$$y_i = w^t x_i$$

We then check the distribution of $\{y_i\}$ (which is the projection of the feature vectors onto the vector w) for both classes to see how well they are separated and if this projection can be used as a classifier.

Fig. 4.9 provides the histogram results for the enhanced mammograms. As can be seen, the two histograms have similar mean and variance and are not well separated. Therefore, for the options we used (enhanced images, segment size, wavelet, etc.) they are not useful for classification.

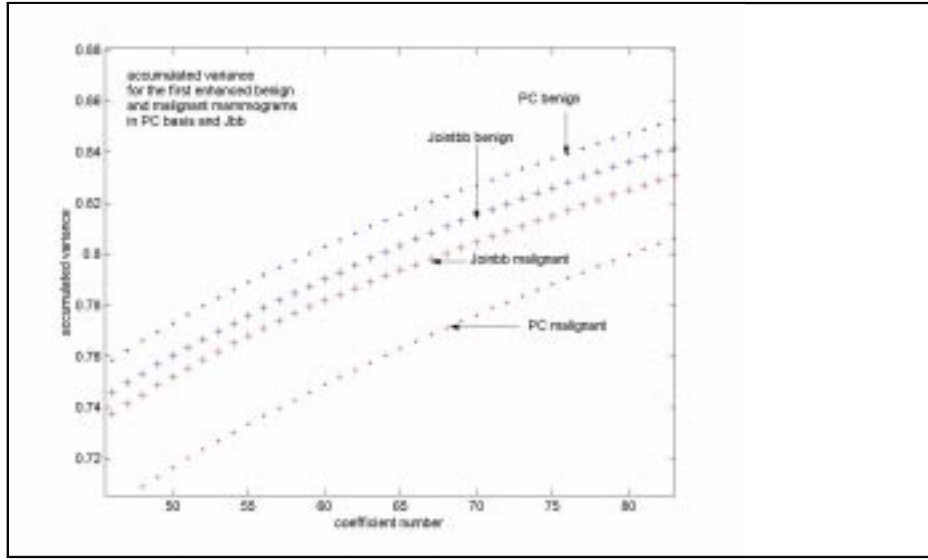


Figure 4.8. Section (magnified) of the accumulated variance in the PC basis and the Joint bb for the first benign and malignant images using segment size of 16x16 pixels.

4.3.1.2 Feature Vectors Based on the Common Joint best basis of Unprocessed Images

We carried out the same experiment using the unprocessed mammogram. Fig. 4.10 shows the histogram of Fisher's Linear Discriminant values for the unprocessed images. There are two ranges of values where there is a significant difference in the distributions: $(-0.2, 0)$ and $(-2, 1)$. In general, the variability in contrast and luminosity in unprocessed images can inflate Fisher's LD values and so Fisher's LD does not provide robust results.

4.3.1.3 Distribution of Wavelet Packet Coefficients in the Common Joint Best Basis

The common Joint best basis was not useful when combined with Fisher's Linear Discriminant. We compute the discriminating power of all coefficients, then select the first two with the largest discriminating power.

The distributions of the first two coefficients with the largest discriminating power are shown in fig. 4.11, where the histogram plots for the enhanced mammograms is shown. The distributions are very similar (when we compare benign to malignant). Both means are zero and the variance for the benign

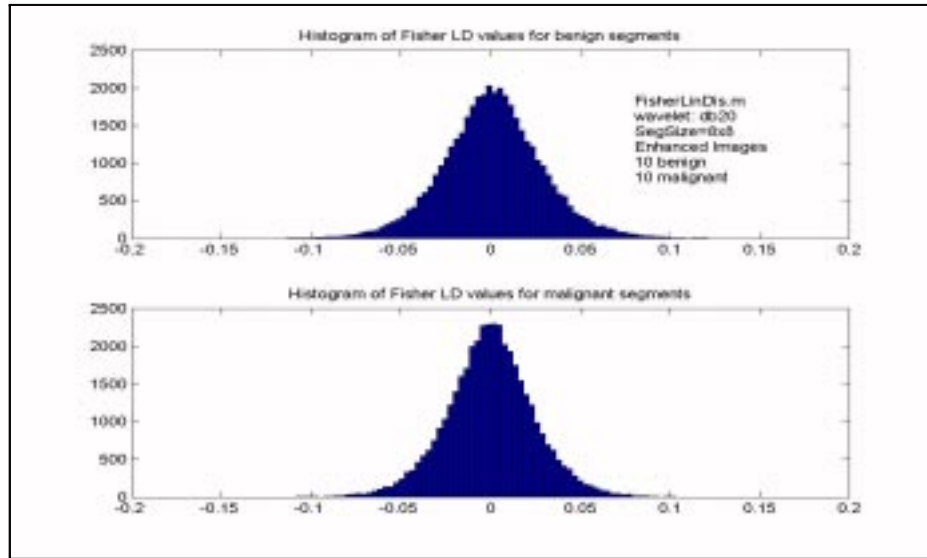


Figure 4.9. Histogram plot of Fisher Linear Discriminant values for enhanced mammograms

image is slightly greater than the variance of the malignant image, but the difference is too small to be used for classification. The distribution is very close to normal and this is due to the normalization effect of the image enhancement we applied to all mammograms.

We carried out the same experiment using the unprocessed mammograms. The results are shown in Fig. 4.12. Note that the distribution of the first coefficient is not normal as was the case for the enhanced mammograms, but that of the second coefficient is close to normal. In general wavelet coefficients with large magnitude are associated with the low frequency and DC components of the signal while the wavelet coefficients with small magnitude are associated with the high frequency components of the signal. Coefficients derived at higher levels of decomposition will have a normal distribution.

Note also that the first coefficient has a distribution with large mean (about 5000 for benign and 10,000 for malignant) while the mean for coefficient 2 is zero and that the variance of the first coefficient is larger than that of the first coefficient.

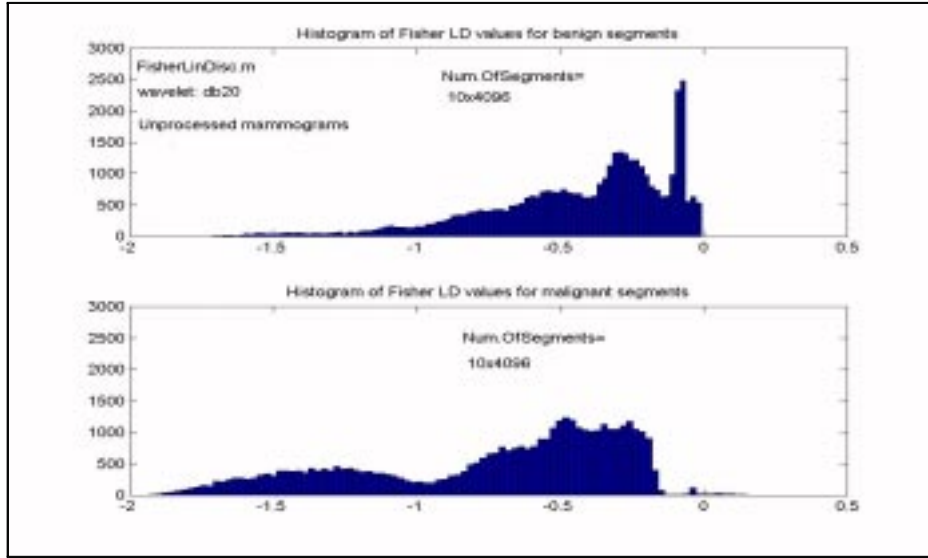


Figure 4.10. Histogram plot of Fisher's Linear Discriminant for the unprocessed images

4.3.2 Classification Results Using a Set of Two Joint Best Bases

In the next experiment we test the approach of using two Joint best bases, one for each class, to be used for classification. We constructed two Joint best bases, one using 4096x10 segments sampled from 10 benign images and the second from 4096x10 segments sampled from 10 malignant images. The two Joint best bases were then used to derive the wavelet packet coefficients for a large ensemble of benign and malignant segments. For each segment, we then computed the distance (the euclidean norm) of its wavelet packet coefficient vector to the average of each class. Figures 4.13 and 4.14 show the scatter plots for the segments used as training data and the results for the test data. We had plotted the two classes in different sets of axes to get a better sense of the separation between the two classes. As can be seen, the locus of the two populations overlap. The distributions of the distance values are similar and they do not provide good discrimination between the classes.

In the next experiment, we reconstruct each segment using 20% of the coefficients (with the largest magnitude) derived from each base. We hoped that the reconstruction error would be smaller for a segment when it was reconstructed using the Joint best basis of its class. Fig. 4.15 is a scatter plot of the ensemble

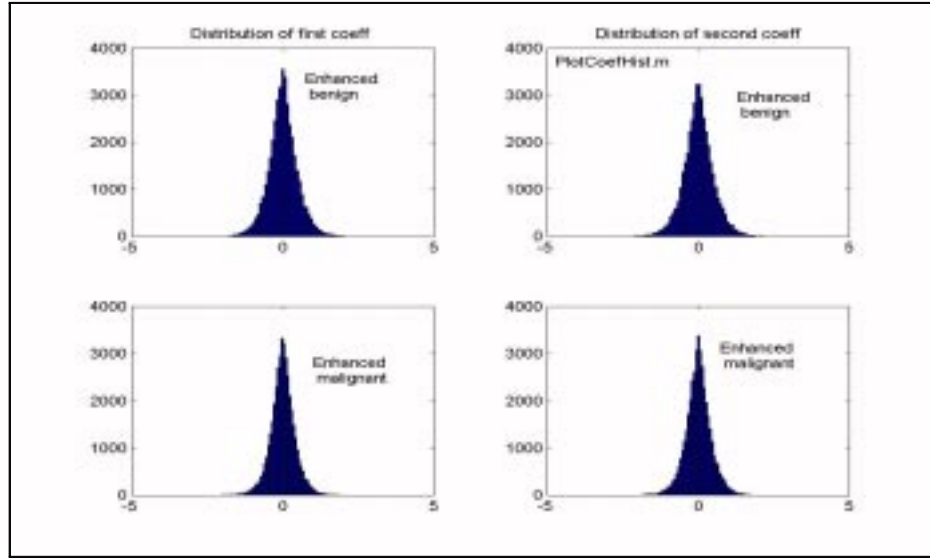


Figure 4.11. Distribution of the first and second wavelet packet coefficients with largest discriminating power for enhanced segments using the common Joint best basis

plotted using the mean squared error (mse) . The segment population of benign and malignant are mixed and there is no separation between the benign and malignant segments. It seems that the variations in the wavelet packet coefficients derived from both Joint best bases take place in many coordinates and there are no dominant coordinates which can be used to discriminate between the two classes.

4.4 Feature Vectors Based on Variance Values

We next present the classification results using as feature vectors the variance values of the wavelet packet coefficients of the ensemble of segments sampled from an image. We use the variance values in the KLT basis and the Joint best basis in two forms: a vector representing their partial accumulated sum (accumulated variance) and a vector representing the partial transform coding gain(TCG).

The following options are used for building the accumulated variance and the (TCG):

Var_8_3_c5

Wavelet used: Coiflet 5

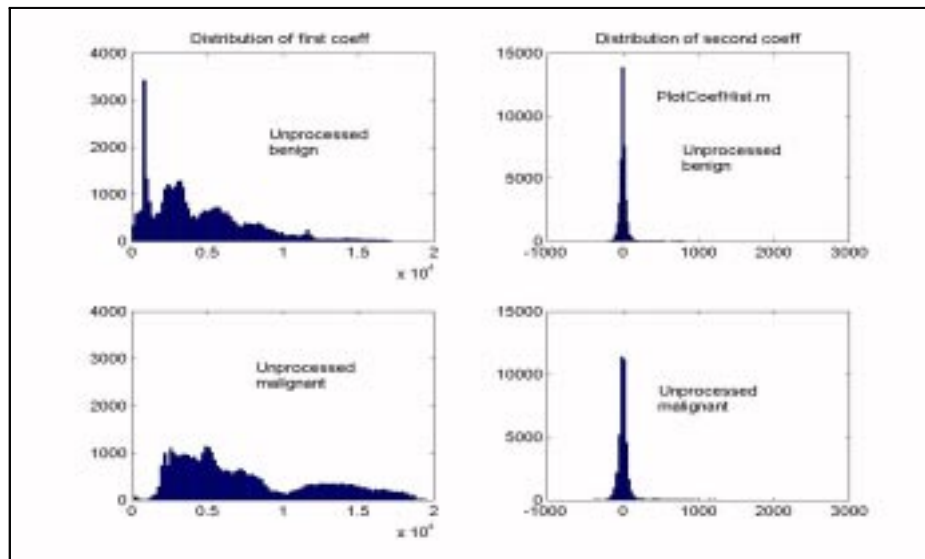


Figure 4.12. Distribution of the first two wavelet packet coefficients with the largest discriminating power for unprocessed images.

Mammogram images used: enhanced

segment size: 8x8

overlapping border: 3 (creates 4096 segment samples for each mammogram).

Joint best basis: derived from the standard deviation of the wavelet packet coefficients.

VarSq_8_3_c5

Wavelet used: Coiflet 5

Mammogram images used: enhanced

segment size: 8x8

overlapping border: 3 (creates 4096 segment samples for each mammogram).

Joint best basis: derived from the squared values of the wavelet packet coefficients.

VarSqUn_8_3_c5

Wavelet used: Coiflet 5

Mammogram images used: Unprocessed

segment size: 8x8

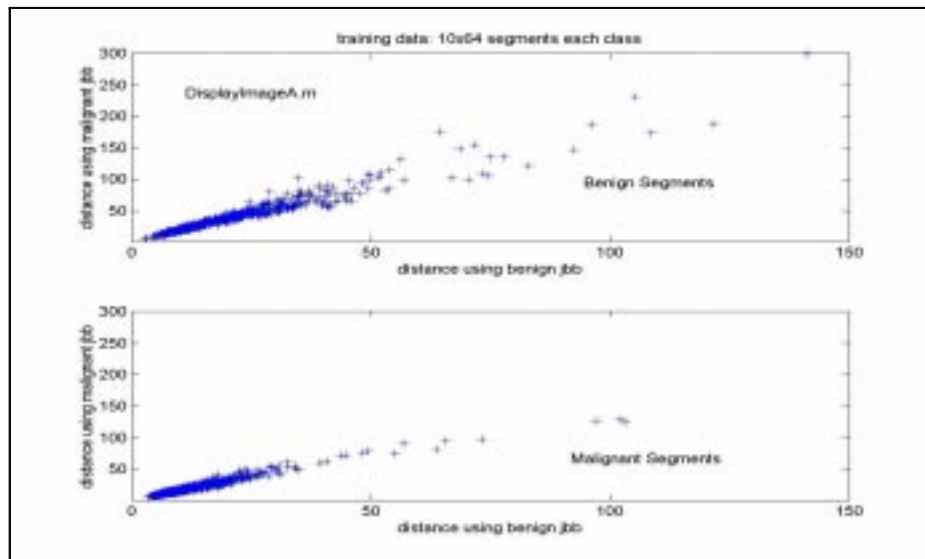


Figure 4.13. Scatter plot of training segments using two Joint best bases

overlapping border: 3 (creates 4096 segment samples for each mammogram).

Joint best basis: derived from the squared values of the wavelet packet coefficients.

VarSqUn_8_3_Haar

Wavelet used: Haar

Mammogram images used: Unprocessed

segment size: 8x8

overlapping border: 3 (creates 4096 segment samples for each mammogram).

Joint best basis: derived from the squared values of the wavelet packet coefficients.

VarSqUn_8_3_db20

Wavelet used: db20

Mammogram images used: Unprocessed

segment size: 8x8

overlapping border: 3 (creates 4096 segment samples for each mammogram).

Joint best basis: derived from the squared values of the wavelet packet coefficients.

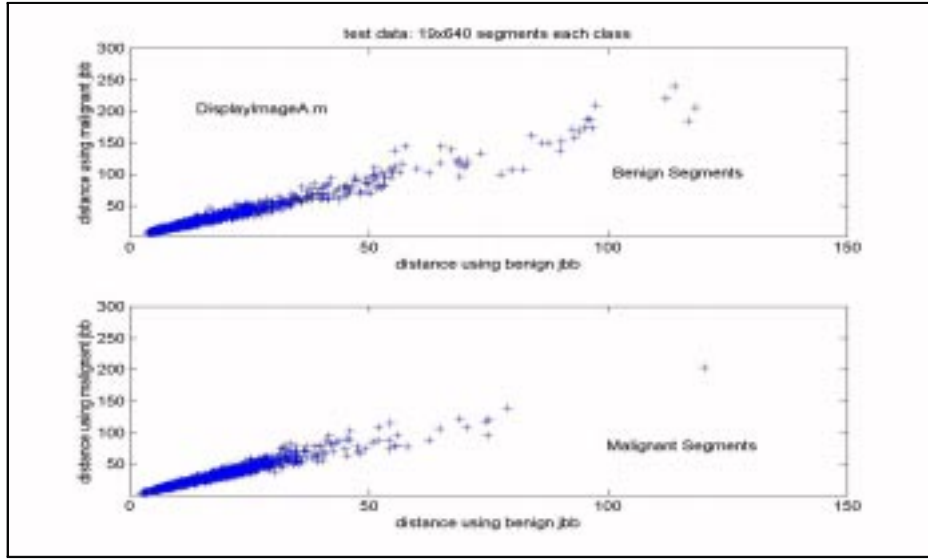


Figure 4.14. Scatter plot of test segments using two Joint best bases

4.4.1 Classification of Mammograms Using the K-nn Algorithm

We will use the K-nn classifier using variance values as feature vectors. First we tested how classification results depend on the number of coefficients we use in the feature vectors. We ran a few experiments using 10,20,30 of the 64 coefficients with the largest discriminating power. We found that in every case except when using the approximate KLT as a feature vector (the results of which will be presented later), there was no improvement in the classification results when more than the first 10 coefficients were used. For illustration we provide figure 4.16 which is a bar plot representing the results of a test with 10 experiments, where each experiment was carried out with 10,20,30 coefficients with the largest discriminating power. As can be seen, the classification results are in average similar for all three cases. We therefore limit the experiments to the first 10 coefficients except for the approximate KLT basis.

To evaluate the performance of each feature vector used for classification, we applied the *Jackknife method*[36]. We created 50 sets of training and test data where we use 70% of the 29 benign and 76 malignant mammograms for training and the rest for testing the classification performance.

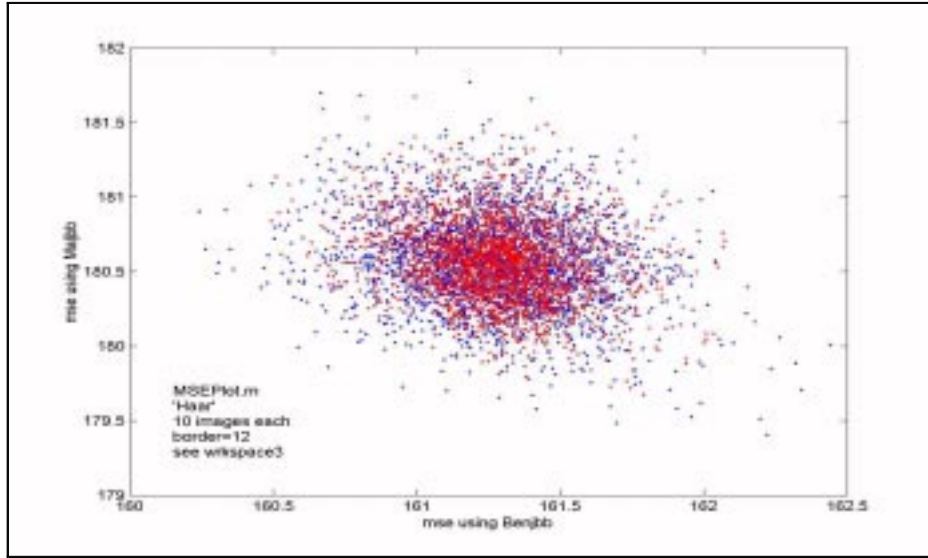


Figure 4.15. Mean Square Error of ensemble of benign and malignant segments, reconstructed using 20% of the coefficient.

The average result over these 50 experiments will be used to evaluate the performance of each feature vector. For each experiment we ran the K-nn algorithm which provides 3 results:

1. The *average error of misclassification* (the average of misclassifying benign as malignant and malignant as benign).
2. The *sensitivity* of the classification, i.e. the percentage of malignant mammograms classified correctly.
3. The *specificity* of the classification, i.e. the percentage of benign mammograms classified correctly.

We performed a set of experiments using various options of the Joint best basis. We derived for each image two feature vectors: the accumulated variance in the Joint best basis(of the image) and the accumulated variance in the KLT basis. For each option we ran a test comprising 50 experiments; then we average the results and used those 3 average values to evaluate the classification performance of the option used.

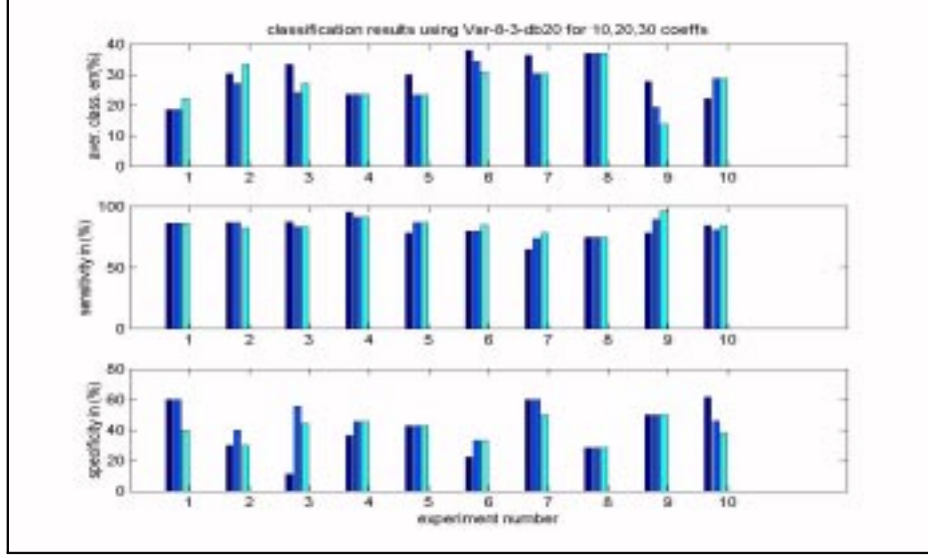


Figure 4.16. Classification results of 10 experiments using the K-nn classifier for 10,20 and 30 coefficients. Top panel-average error, mid panel-sensitivity, bottom panel-specificity.

4.4.2 Feature Vectors Based on the Transform Coding Gain(TCG)

As was described in chapter 3, TCG is a measure frequently used for the compression property of a transformation [26]. The accumulation of variance and the TCG are related. We would use both as feature vectors since the different mathematical operators involved in each may capture a 'signature' of the image that may be useful for discrimination.

In the next experiment we use the TCG of KLT basis and the Joint basis as feature vectors combined with the K-nn classifier.

We defined the n^{th} partial transform coding gain of the KLT basis, $G_{TC(KLT)}^n$:

$$G_{TC(KLT)}^n = \frac{\frac{1}{n} \sum_{i=1}^n \sigma_i^2}{\left(\prod_{i=1}^n \sigma_i^2\right)^{\frac{1}{n}}} \quad n = 1, 2, \dots, d$$

and the transform coding gain vector of the KLT basis, $V_{TCG(KLT)}$:

$$V_{TCG(KLT)} = \{G_{TC(KLT)}^1, G_{TC(KLT)}^2, \dots, G_{TC(KLT)}^d\}$$

Similarly we defined the n^{th} *partial transform coding gain* for the Joint best basis, $G_{TC}^n(jointbb)$

to be:

$$G_{TC}^n(jointbb) = \frac{\frac{1}{n} \sum_{i=1}^n \sigma_i^2}{(\prod_{i=1}^n \sigma_i^2)^{\frac{1}{n}}} \quad n = 1, 2, \dots, d$$

and the *transform coding gain vector* for the Joint best basis of the segment population, $V_{TCG}(jointbb)$:

$$V_{TCG}(jointbb) = \{G_{TC}^1(jointbb), G_{TC}^2(jointbb), \dots, G_{TC}^d(jointbb)\}$$

We represent each image by these two feature vectors where σ_i^2 is the variance of the wavelet packet coefficients of the ensemble of segments (sampled from an image) along coordinate i . These two vectors will be used as input features to the K-nn classifier. As before, the rational is that the KLT basis and Joint best basis are bases in which the distribution of the variance of the segment population carry a signature of the image. We hope that the signature of images within each class will have small variability and at the same time the signatures in both classes will have a significant difference that can be used for classification.

Fig. 4.17 shows a plot of the discriminating power of a feature vector based on the TCG in the KLT and the Joint best basis for the option Var-8-3-db20. Note as in the case of the accumulated variance, that the two vectors do not achieve their peak values for the same coefficient.

To get a robust evaluation of the performance of these feature vectors, we applied the K-nn classifier as before. The average results for a few of the options are provided in the next table:

Classification results using TCG in the KLT and JBB as feature vectors			
	average error	sensitivity	specificity
Var_8_3_c5	35.0480	76.6081	39.3409
VarSq_8_3_c5	35.0776	75.8895	37.5046
Var_8_3_Haar	38.4844	72.0439	34.9556
Var_8_3_db20	38.0494	75.6751	31.5379

The best results we get for the TCG as a feature vector is for the option Var_8_3_c5, with 35% average error, 76.6% sensitivity and 39.3% specificity. These results are close to the average performance of radiologists.

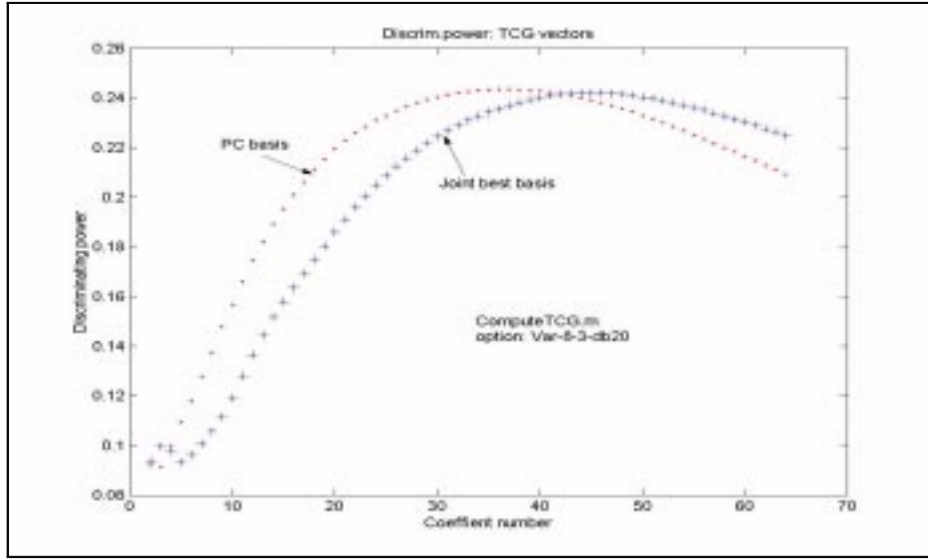


Figure 4.17. Discriminating power plot using the Transform Coding Gain as feature vectors.

According to studies [34], the average performance of radiologist is 80% for sensitivity and 20% for specificity.. The results of our next experiments are better and they will be presented in the next section.

4.4.3 Feature vectors based on the accumulated variance

We represent each mammogram by two vectors: the accumulated variance using the KLT basis and the Joint best basis (total of 128 coefficients). We will later apply the K-nn for classification and choose only 64 of the 128 coefficients with the largest discriminating power.

To examine the discrimination power of the feature vectors based on the accumulated variance in the two bases, Figures 4.18 and 4.19 provide the discriminating power of each coefficient for the first two options described above. Notice that the two feature vectors get their discriminating power peak value at different coefficient number (for the first option, the graph of the KLT basis gets its peak value for coefficient number 2 while the Joint best basis gets its peak value for coefficient number 8). The K-nn classifier will sort the coefficients in both vectors according to their power of discrimination and will use the coefficients with the largest values.

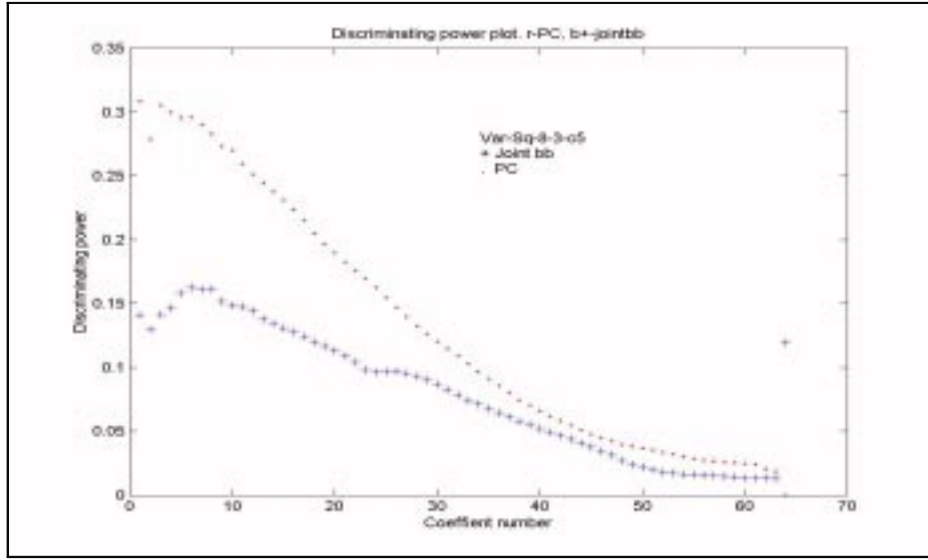


Figure 4.18. Discriminating power plot of the accumulated variance in the PC basis and the Joint best basis for the first option.

For illustration and to get some visual evaluation of the classification performance of the two feature vectors, we present in Figures 4.20 and 4.21 a scatter plot of all mammograms (29 benign and 76 malignant) for the first two options. For the first option we use the accumulated variance of the first 6 coefficients and for the second option we use the first 42 coefficients.

Notice in the scatter plots that the error for both the KLT basis and the Joint best basis are lower for the malignant class than for the benign. This is probably because there is some structure in the malignant class so that the Joint best basis of the malignant segments sampled from a malignant mammogram is in general a "better basis" (in terms of its transform coding gain) than a Joint best basis for a benign mammogram. This is also true for the error in the KLT basis.

As we mentioned, to get the full discriminating power of these feature vectors we will use the K-nn algorithm which selects a subset of the coefficients with the greatest discriminating power.

We provide the full test results of 50 experiments for the feature vector based on the option Var_8_3_db20 in Figure 4.22.

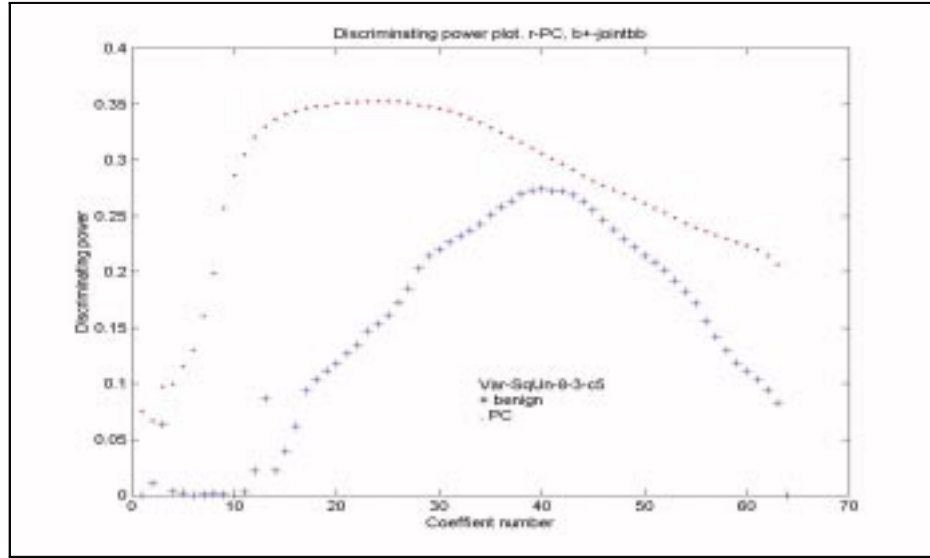


Figure 4.19. Discriminating power plot of the accumulated variance in the PC basis and the Joint best basis for the second option

To compare the performance of each option used for classification, we use the average results to compare the classification performance under different conditions; e.g. different wavelets, enhanced or unprocessed image.

The next table summarizes the average results for each option:

Classification results using the accum. var. in the KLT basis and JBB as feature vectors

	average error	sensitivity	specificity
Var_8_3_c5	29.8397	79.1575	47.7302
VarSq_8_3_c5	30.8052	80.4606	40.1165
Var_8_3_Haar	28.4920	82.6875	37.1411
Var_8_3_db20	29.4324	80.5776	48.4636
VarSqUn_8_3_c5	37.5597	74.7326	29.2680
VarSqUn_8_3_Haar	36.9473	72.8653	40.4315
VarSqUnV0_8_3_c5	31.2609	76.6828	51.4574

4.4.3.1 Evaluation of the results

The results of the first 4 options (which use enhanced images) have a better sensitivity than the those of the last 3 options (which use unprocessed images): 79-82% compared to 72-76% with a better specificity, so that the enhancement of the images does provide improvement on the classification of the

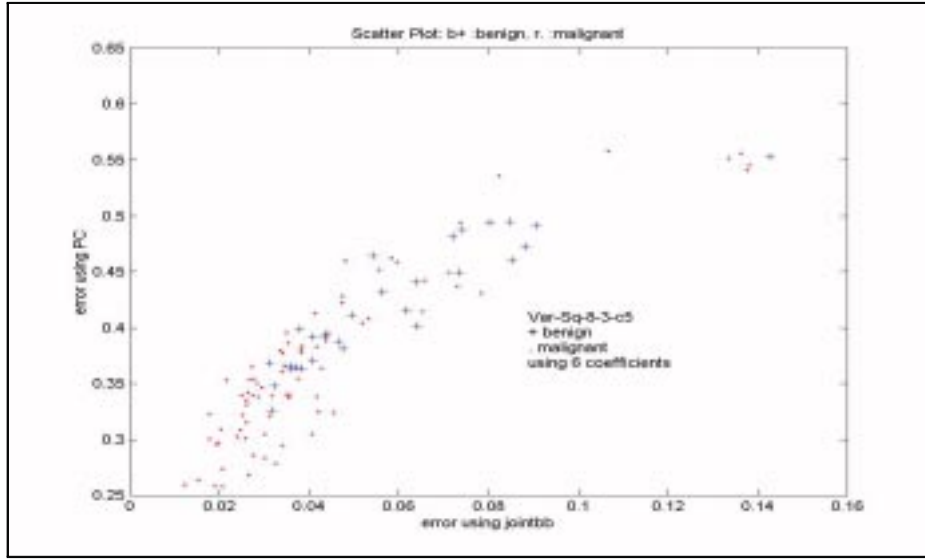


Figure 4.20. Scatter plot of mammograms using option 1 with 6 coefficients of the accumulated variance mammograms. Among all options, Var_8_3_db20 did the best (the Joint best basis is derived from the standard deviation of the wavelet packet coefficients and the wavelet used is the 20- tap Daubechies filter) with an average error of 29.4%, an average sensitivity of 80.5% and an average specificity of 48.4%. We used the db20 filter as this filter was reported to have a good performance in detecting calcification points [36].

According to studies [34], only 10-30% of women who undergo a biopsy have cancer (an average specificity of 20%) while 10-30% of malignant mammograms are missed by radiologists (average sensitivity of 80%). Comparing the average performance of radiologists (80% sensitivity and 20% of specificity) to the performance of the Var_8_3_db20 option, the sensitivity of our classification is approximately the same while the specificity is better (48.4% compare to 20%).

We emphasize again that the mammograms used in these experiments are from a general screening of women population. Therefore they are harder images (in terms of discrimination) when compared to mammograms taken by women due to some pathological indications (e.g. pain, lumps in the breast, asymmetry in the breasts). For comparison, in [34] the images used are radiographs of biopsy specimens. In biopsy specimens, scatter radiation recorded on films is reduced because there is less underlying tissue

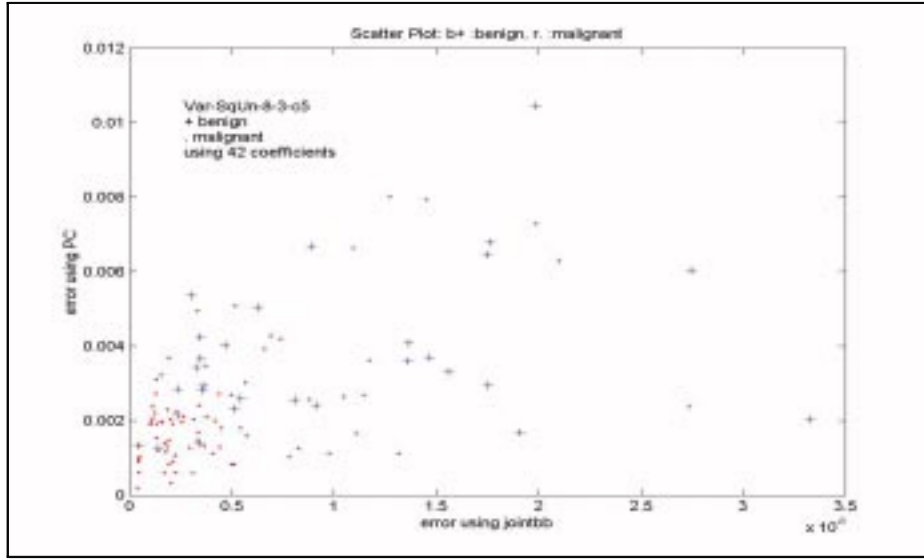


Figure 4.21. Scatter plot of mammograms using option 2 with 42 coefficients of the accumulated variance.

around microcalcifications than in normal mammograms. Therefore, microcalcifications in radiographs of biopsy specimens are generally more clearly represented than those in regular mammograms[34].

These results suggest that the feature vector we use combined with the K-nn classifier are useful in early detection of breast cancer.

4.4.4 Classification Results Using the Approximate KLT

Next we used the accumulated variance in the approximate KLT basis and the Joint best basis as a feature vectors. Each of this bases was computed for each image. We then applied the K-nn classifier as before to these feature vectors, choosing randomly 70% of the data as training data and the rest as test data. Figure 4.23 shows the bar plot for the results of 50 experiments.

We use the average of these 50 experiments to evaluate the combined accumulated variance in the approximate KLT basis and the Joint best basis as feature vectors. The average results for this set of 50 experiments are:

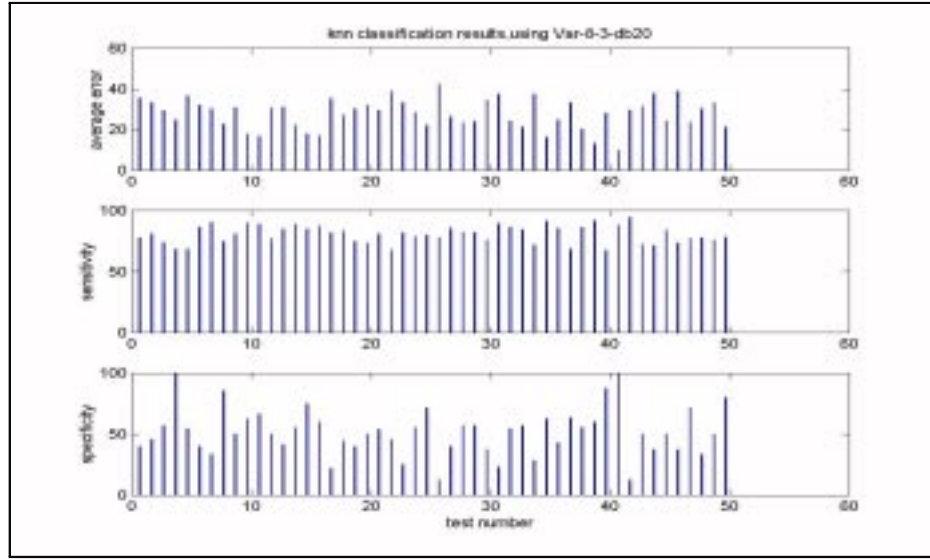


Figure 4.22. Classification results of 50 experiments using 10 coefficients and the Knn classifier

Classification results using the approximate KLT basis and the Jbb

average error	sensitivity	specificity
25.2728	83.2612	55.1209

This results are better than the best results we achieved using the accumulation of variance in the KLT basis and the Joint best basis (option:Var_8_3_db20). We provide the results for both options for comparison:

Comparison of classification results using the approximate KLT and the KLT as feature vectors			
	average error	sensitivity	specificity
Var_8_3_db20 using accum.var. in KLT basis and the Joint bb	29.4324	80.5776	48.4636
using accum.var. in approx. KLT basis and the Jont bb	25.2728	83.2612	55.1209

The results using the approximate KLT basis are better. The average error reduced to 25.27%, the sensitivity increased to 83.26% (about 3.26% better than the average sensitivity of radiologists) and the specificity increased to 55.12% (about 35% better than the average specificity of radiologists).

The approximate KLT and the Joint best basis combined with the K-nn classifier provide the best classification results in this study, when compared to all other feature vectors we used.

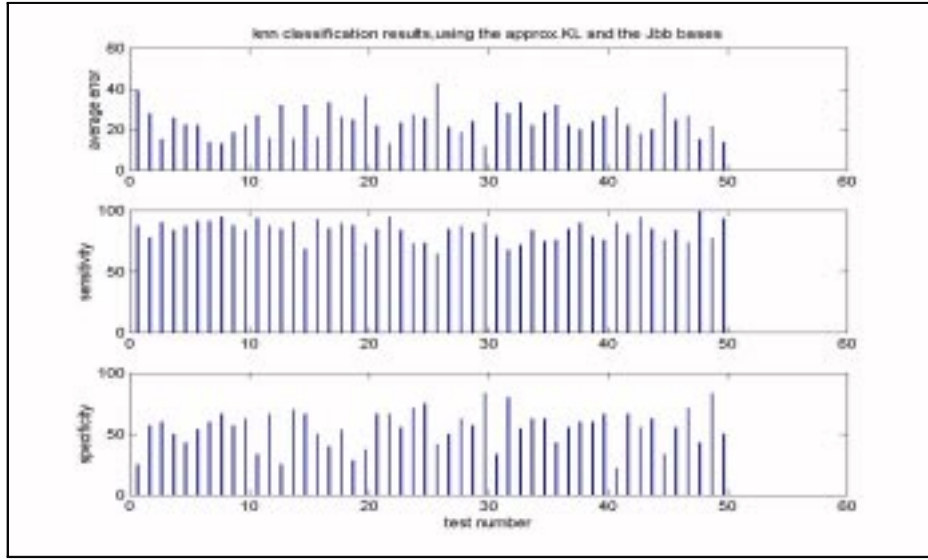


Figure 4.23. Classification results of 50 tests using the Approximate KL Transform as feature vector

4.4.5 Biplot using the approximate KLT basis and the Joint best basis

The approximate KLT basis provided the best classification results when we used all the 64 coefficients in the accumulated variance. In certain cases the first coefficients (with the largest discriminating power) carry enough discriminating power and a biplot of the data using only two coefficients will result in two clusters. Figure 4.24 is a biplot of all images using the first coefficient in the approximate KLT basis and the first coefficient in the Joint best basis (both with the largest magnitude). As can be seen, there is no good separation of the benign and malignant mammograms, implying the first coordinates are not sufficient to discriminate between the two classes.

4.4.5.1 Comparing variance values and accumulated variance values as feature vectors

In the next experiment we provide some results that justify the use of the accumulated variance as a feature vector. We compared the performance of:

The variance values directly as feature vectors.

The accumulated variance as feature vectors.

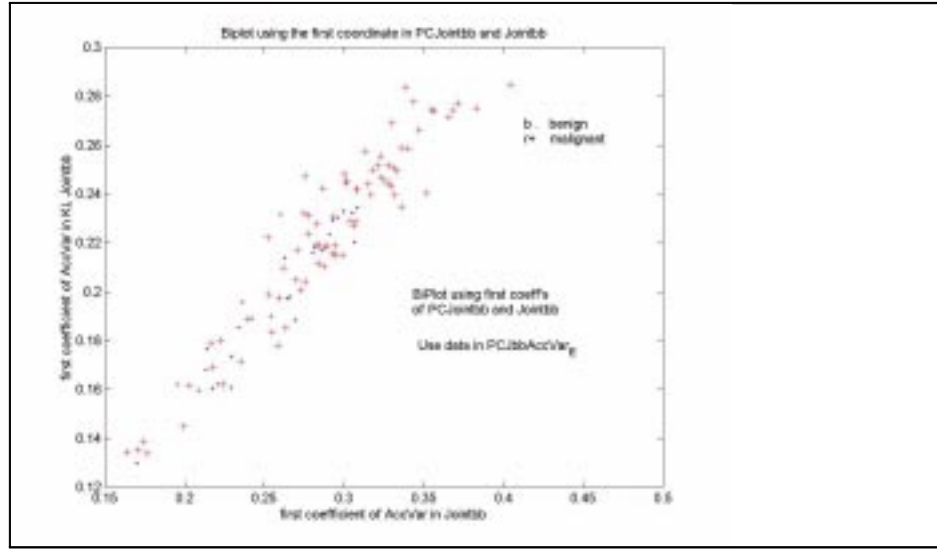


Figure 4.24. Biplot of mammograms using the first coefficients in the approximate KL basis and the Joint best basis

The next table provides the average results of 50 experiments for each option . The accumulated variance as feature vector provides better classification results than the variance values as feature vectors. The average error is lower by approximately 5%, the sensitivity higher by 6% and the specificity higher by almost 6.5%.

Comparison of classification results using the approximate KLT and the KLT as feature vectors

	average error	sensitivity	specificity
using var. values in approximate KLT	30.6165	77.3437	48.5990
using accum.var. in approximate KLT	25.2728	83.2612	55.1209

4.5 Summary

In this chapter we have shown that the compression properties represented by the accumulated variance in different data representations of the mammograms provide a good indicator for discrimination. Among all feature vectors used in this study, the feature vectors based on the accumulated variance in the approximate KLT basis and the Joint best basis derived for each image, provided the best discriminant between benign and malignant mammograms. These results provide a classification performance better

than the average performance of radiologists: 83% sensitivity compared to 80% and 55% specificity compared to 20%)[34].

The best results were achieved using the enhanced mammograms, justifying the image enhancement step we applied to the mammograms in the data base and using the wavelet db20, compatible with previous research that suggested this filter is best for detecting calcification points [36].

The data base used in this study is among the most difficult data sets as it contains mammograms of randomly chosen women (with clinical indications of breast cancer); namely it contains mammograms where the possible malignant tumors are very small and thus very difficult to detect. Radiologists performance at this stage of the disease is significantly lower than the performance of more advanced stages. Detection of breast cancer at early stage enables treatment which is much more effective, less invasive and inexpensive.

Chapter 5

Combining the Hotelling Transform In Image Query

1. Introduction

In the "content based" image query, also referred to as "query by example", "similarity retrieval" or "sketch retrieval", the query image is provided by the user either as a sketch of the object, as the output of a scanner or a video camera. Some of the difficulties associated with content based image query are described in [1], e.g. significant differences between the "query image" and the "target image", artifacts and poor resolution of the query image make a straightforward comparison of images using L^1 and L^2 metrics not effective.

In [2] a new strategy is suggested based on wavelet decomposition of the query image and the database images combined with a metric which is designed to be insensitive to small differences in the query process. This approach is found to be fast and overcomes the above mentioned problems.

Wavelet coefficients of an image may vary strongly when the image is displaced or rotated (unlike color histogram of an image which is invariant under displacement and rotation). Although the metric suggested in [2] is more robust to these errors when compared to the L^1 and L^2 metric (but worse when compared to the metric based on color histograms), still the error is significant.

In this paper we provide some experimental data on the sensitivity of the wavelet coefficients to displacement and rotation in the context of the standard characters and suggest an integration of the Hotelling transform to improve on this sensitivity. We also provide some experimental data on the distribution of the largest wavelet coefficients at different levels of the wavelet decomposition and discuss some questions relevant to the approach of using a set of the largest wavelet coefficients for image query.

Overview of this paper

Section 2 describes the content based image query using wavelet decomposition as described in [1] and [2], including the L^q metric used to compare the wavelet coefficients of the query image and the target image. In section 3 the Hotelling transform is described including its algebra. Section 4 describes

some experimental results using MATLAB regarding the sensitivity of the L^q metric to rotation and displacement of an image. Section 5 will describe the use of the Hotelling transform to improve the image query process. In section 6 we discuss some questions regarding the set of the largest wavelet coefficients as representing an image: its size, its distribution at different levels of the wavelet decomposition and the L^q metric used to compare two such sets.

2. Content based image query using wavelet decomposition

Some of the previous approaches to content based image query, include the use of certain properties of the images or a combination of these properties. For example, the user may specify a color combination (color histogram), a texture, geometrical features(e.g. edges, shapes or major axis orientation) or a rough sketch of the image. The multiresolution approach suggested in [2] appears to have a success rate at least as good as that of other systems that work from a simple user sketch. This approach has some advantages when compared to the previous ones. First, It decouples the resolution of the query image and the target image and hence the query image can be specified at any resolution. Second, the performance (running time) is independent of the resolution of the database images(only the set containing the largest wavelet coefficients is used in the comparison regardless of the resolution of the images) and third, it provides a simple algorithm and a compact code in the implementation.

As described in [2], a wavelet decomposition is applied to the database images(color images, of size 128x128) , using Haar wavelets (simple to compute) and standard decomposition. Then, all but the 60 (experimental result) largest (in magnitude) coefficients are truncated for each color channel . These remaining 60 coefficients are then quantized to two levels only, +1 and -1 for positive and negative coefficients respectively. The metric developed (denoted by L^q), compares only the indices of these 60 largest, quantized coefficients of the query image with those of the data base images, and the scores of this comparisons are used to pull the 20 best matches from the database. The truncation and quantization are significant in this approach, since they reduce the search time and the storage requirements, but more importantly they improve the performance of the L^q metric used in comparing the query image to the target. This is because the truncation and quantization make the L^q metric tolerant to small differences

between the query image and the target while giving more weight to the significant features that are closely matched. When compared to the L^1 and L^2 metrics (which take in account all the 128^2 wavelet coefficients of the images), and the L^c metric, the metric that uses color histograms, the L^q metric has a better success rate.

As to the robustness of the multiresolution approach to distortion due to image rotation and displacement of the query image, when compared to the L^1 and L^2 metrics, the L^q metric has a better success rate, but when compared to the L^c metric, its performance is worse (since color histograms are not sensitive to rotation and translation).

3. The Hotelling Transform and the principal axis of an image

The Hotelling transform is a linear transformation of a set of n dimensional vectors that decorrelates the n coordinates. When applied to an 2-dimensional image, the transformed image will be aligned along its principal axis.

The Hotelling transform can be used to improve the robustness to distortion of the image due to rotation and displacement. The improvement though, would increase the running time of the image query.

This transformation is a combination of displacement and rotation of the object. The centroid of the image(the average of the x and y coordinates of all pixels) is shifted to the origin and the image is rotated by an angle that minimizes its moment of inertia[4]. Geometrically the transformed image will be oriented in the direction in which it seems to be the most 'elongated'.

In image recognition this transformation is helpful, since the identity of the object is not known and aligning the image with its principal axis can help remove the effects of translation and rotation in the analysis.

First we describe the Hotelling transform. Given a set k of n -dimensional column vectors: X_1, X_2, \dots, X_k , the covariance matrix of the vector population is given by:

$$C_x = E[(X - m_x)(X - m_x)^T]$$

where m_x is the mean vector of the population.

C_x is a $n \times n$ real symmetric matrix whose c_{ij} entry equals the covariance between the i^{th} and the j^{th} coordinates of the vector population. If $c_{ij} = 0$, then the i^{th} and the j^{th} coordinates are decorrelated. When $c_{ij} > 0$ or $c_{ij} < 0$ then there is a positive or a negative correlation between the i^{th} and the j^{th} coordinates, respectively.

The Hotelling transform maps the given vector population, X , into a vector population Y (that consists of k , n -dimensional vectors) such that $m_y = 0$ and C_y , the covariance of the new vector population is a $n \times n$ diagonal matrix and therefore the i^{th} and the j^{th} coordinates of the new vector population are decorrelated for all $i \neq j$. To see the relevance of the Hotelling transform to aligning 2-dimensional image with its principal axis, consider an image composed of $m \times n$ pixels. If the (x, y) coordinates of each pixel is presented as a 2-dimensional vector, the image then is represented by a set of $m \times n$, 2-dimensional, vectors. When the Hotelling transform is applied to this vector population, the mean vector of the new vector population, $m_y = 0$, and its covariance will be diagonal, which means that the x and y coordinates are decorrelated.

Geometrically this would mean that the centroid of the original image was shifted to the origin and the image was rotated so that its principle axis is aligned with the X -axis.

Algebra of the Hotelling transform

Let X be a set of k , n -dimensional column vectors: X_1, X_2, \dots, X_k

The mean vector of the population, m_x , is given by:

$$m_x = \frac{1}{K} \sum_{i=1}^K X_i$$

and the covariance matrix of the vector population, X , is:

$$\begin{aligned} C_x &= E[(X - m_x)(X - m_x)^T] = \\ &= E(XX^T) - E(Xm_x^T) - E(m_x X^T) + E(m_x m_x^T) = \\ &= E(XX^T) - m_x m_x^T - m_x m_x^T + m_x m_x^T = \\ &= \frac{1}{K} \sum_{i=1}^K X_i X_i^T - m_x m_x^T \end{aligned}$$

Now, C_x is a $n \times n$ real and symmetric matrix (because it is the sum of the products of vectors by their transpose and the entries of the vectors, which are the pixel coordinates, are real values); therefore it has n real eigen values and n real orthogonal eigen vectors[5].

Let $\lambda_1, \lambda_2, \dots, \lambda_n$ be the n eigen values with $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$ and V_1, V_2, \dots, V_n be the corresponding set of orthonormal eigen vectors, where V_1 corresponds to λ_1 , the largest eigen value.

Define the matrix A whose rows are the n eigen vectors V_1, V_2, \dots, V_n :

$$A = \begin{bmatrix} V_1 & - & - & - & - & > \\ V_2 & - & - & - & - & > \\ & & & & \cdot & \\ & & & & \cdot & \\ V_n & - & - & - & - & > \end{bmatrix}$$

The Hotelling transform is then the linear transformation given by:

$$Y_i = A(X_i - m_x) \quad i = 1, 2, \dots, k$$

The mean of the vector population Y , is 0, since

$$\begin{aligned} m_y &= \frac{1}{K} \sum_{i=1}^K Y_i = \\ &= \frac{1}{K} \sum_{i=1}^K A(X_i - m_x) = \\ &= A \frac{1}{K} \sum_{i=1}^K (X_i - m_x) = 0 \end{aligned}$$

and the covariance matrix of the vector population Y is diagonal:

$$\begin{aligned} C_y &= E[(Y - m_y)(Y - m_y)^T] = \\ &= \frac{1}{K} \sum_{i=1}^K Y_i Y_i^T - m_y m_y^T = \\ &= \frac{1}{K} \sum_{i=1}^K Y_i Y_i^T = \\ &= \frac{1}{K} \sum_{i=1}^K [A(X_i - m_x)][A(X_i - m_x)]^T = \\ &= \frac{1}{K} \sum_{i=1}^K A(X_i - m_x)(X_i - m_x)^T A^T = \end{aligned}$$

$$\begin{aligned}
&= A \left[\frac{1}{K} \sum_{i=1}^K (X_i - m_x)(X_i - m_x)^T \right] A^T = \\
&= AC_x A^T = \\
&= \begin{bmatrix} V_1 \longrightarrow \\ V_2 \longrightarrow \\ \vdots \\ V_n \longrightarrow \end{bmatrix} C_x \begin{bmatrix} V_1 & V_2 & \cdots & V_n \\ \downarrow & \downarrow & \downarrow & \downarrow \end{bmatrix} = \\
&= \begin{bmatrix} V_1 \longrightarrow \\ V_2 \longrightarrow \\ \vdots \\ V_n \longrightarrow \end{bmatrix} \begin{bmatrix} \lambda_1 V_1 & \lambda_2 V_2 & \cdots & \lambda_n V_n \\ \downarrow & \downarrow & \downarrow & \downarrow \end{bmatrix} = \\
&= \begin{bmatrix} \lambda_1 & & & \circ \\ & \lambda_2 & & \\ & & \ddots & \\ \circ & & & \lambda_n \end{bmatrix}
\end{aligned}$$

(where the last equality is because the eigen vectors are orthonormal).

4. Sensitivity of the Wavelet Transform to rotation and displacement of an image

Next we present some experimental data on the sensitivity of the wavelet transform to rotation and displacement in the context of the standard characters.

Let an image and its rotated version be represented by the matrices X and X_R respectively and the wavelet transform be represented by the matrix W . Since the wavelet basis is an orthonormal basis, it preserves the L^2 norm [8] and so the relative error when comparing X and X_R , i.e. $\frac{\|X - X_R\|}{\|X\|}$, equals the relative error when comparing their wavelet coefficients, i.e. $\frac{\|WX - WX_R\|}{\|WX\|}$.

But when the wavelet coefficients are truncated and quantized, the relative error in comparing only a limited set of the coefficients (largest in magnitude) depends on the distribution of the coefficient values. The wavelet transform will contain many coefficients whose magnitude is small for images with high degree of regularity. Following [2], we compare the 60 largest coefficients and present the error as the number of mismatches.

The experimental results were produced using MATLAB Wavelet Toolbox and the Image Processing Toolbox. The images contain monochromatic bitmap alphabet characters, standard upper case New

Courier, size 72, created using PAINT-BRUSH. Each character was first aligned so that its centroid was at the center of a 228x228 image.

The results given are relevant in the context of the standard alphabet characters and they may vary for different types of database images, hence for a specific application this experiment should be duplicated to get a better representation of the errors.

The first test provides the data on the sensitivity of the largest 60 wavelet coefficients to rotation of the image. To avoid the error resulting from the mismatch of the edges of the rotated image when compared to the original, the central (128x128) portion (which includes the character) of the 228x228 image was rotated and the largest 60 wavelets coefficients were quantized to the values ± 1 and compared to the largest 60 wavelet coefficients of the center portion of the unrotated character.

We use the same range and resolution as in [2] to present the errors associated with rotation and translation of the image.

Fig. 1 shows the relative error in % (i.e., number of mismatches *100%/60) for some of the standard alphabet characters. Results are presented for angle of rotation in the range 0 to 45 degrees in steps of 5 degrees. The relative error is closely linear with respect to the angle of rotation. Note that the relative error for the characters 'C' and 'D' which have radial symmetry, is less than the error for the characters 'B', 'A' and 'Z'. The errors for the latter characters are significantly high and are in the range of 40% to 50% for a rotation of 10 degrees.

In the second test we checked the effect of displacement on the largest wavelet coefficients. The character was displaced by a certain percentage of the width of the character (0 to 50%, in steps of 5%) and the 60 largest wavelet coefficients were compared to those of the original image. Fig. 2 shows the results for the same standard alphabet characters. While the relative error for the different characters is closely the same, note that the error in general is very significant due to displacement for all 5 characters. For example, a displacement of the image by 0.1 of the width of the image results in a relative error in the range of 75%-85% . The error due to displacement will be removed by the Hotelling transform since the transformed image will be aligned with its centroid at the origin.

Fig. 3 and 4 provide some results on the error due to rotation and translation for a few different wavelet bases. The 'haar' wavelet is more sensitive to rotation and displacement when compared to 'db2', 'db4' and 'coif' wavelets. This is because the latter wavelets have a wider support and use a larger neighborhood resulting in stronger averaging effect which make them less sensitive to rotation and displacement. On the other hand, this means that they would have worse performance in discriminating between different images (compared to 'haar').

5. Combining the Hotelling transform in the query process

As was mentioned before, the error due to displacement will be eliminated by the Hotelling transform since the image centroid will be aligned with the origin. The problem is not simple when it comes to the error due to rotation. The angle of orientation of an image, depends on the geometry of the image. Thus some images may have a principle axis which may be sensitive to slight variations in the geometry of the image and others may not. For example, objects with a geometry that strongly resembles a rectangular would have a principal axis not sensitive to small distortions, while a geometry that resembles a square or circle would be sensitive to small distortions. Because of that, the Hotelling transform cannot be integrated in a straightforward manner and rather than improving the query, it may make it worse, since these objects may be rotated by an unpredictable and undesirable angle.

To overcome this problem, once the query image is given and its principal axis computed by the Hotelling transform, the angle of rotation that corresponds to the principal axis can be checked. If the value is within some permissible window(which will depend on the application) the transformed image(i.e. the centered and rotated image according to its principal axis) can be used as the query image. Otherwise, we form a set of images, (which we call an image query set) that will include the original query (with its centroid aligned at the origin) and a few rotated (both clockwise and anti clockwise) versions of the image query. This set then will be used in the query process. Rather than comparing a single query image, each of the images in the query set will be compared to the targets in the database, and the score can be a weighted sum of these comparisons(or alternatively, the score can be that of the best match among the rotated versions).

As to what is the permissible angle and how many images the query set should include, this would depend on the application, the type of images, the desirable search speed and the level of error since there is a trade off between the query time and the success rate. Data that provide the relative error as a function of rotation can help in determining the width of the permissible window and the number of images in the image query set.

In the implementation of the Hotelling transform, the eigen vector corresponding to the largest eigen value of the covariance matrix provides the angle that the image has to be rotated so that its principal axis would be aligned with the X-axis.

Table 1 provides some experimental results illustrating the improvement that can be achieved by using the Hotelling transform to align an image before its comparison. The Hotelling transform was implemented in MATLAB using the 'Haar' wavelet, and the non-standard decomposition. The table shows first the error in comparing a hand sketched query character with its standard counterpart and the error in comparing the transformed same hand sketched character and its counterpart standard character. In the implementation of the Hotelling transform, once the image is centered, the rotation angle provided by the Hotelling transform was checked. If the angle was less than 10 degrees, the image would be rotated otherwise it will only be shifted so that its centroid is at the origin. Results are presented for a few different alphabets with a distinct principal axis.

Table1 : The relative error (%) between the largest 60 coeffs of a hand sketched query character and its counterpart standard character with and without the Hotelling transform

character	<i>Relative Error(%)</i>	
	H.T. applied to query char.	Not applied
A	45	91
I	85	98
J	85	98
M	60	81
S	86	86

6. Optimal set of the largest wavelet coefficients and their distribution at different levels of the wavelet decomposition

Next we present some questions relevant to the approach that uses the set of the largest wavelet coefficients for image query.

The first question is how reliable can a set of largest wavelet coefficients be as a discriminating tool in image query. Clearly the answer depends on the type of images used in the data base. Can certain features in the images comprising the data base(e.g. edges, contour lines, the gradient of an image) can be captured by the set of the largest wavelet coefficients so that they can be used as a discriminating tool in image query? As an experimental approach one may take a representative sample of the data base, then compute a matrix which provides the error in all pairwise comparisons in this set, then examine the data to see if there is a significant difference between the error in comparing images that closely resemble each other and the error in comparing images that differ significantly. As an example, table 2 provides such data for some of the standard alphabet characters using the 60 largest coefficients.

table 2 : The relative error(%) in comparing (pairwise) some of the alphabet characters

	A	B	C	D	E	F	G	X	Y	Z
A	0	65	68	63	68	73	57	58	83	73
B		0	68	48	58	78	58	60	73	42
C			0	50	73	72	52	75	70	73
D				0	65	73	65	68	75	65
E					0	77	68	62	80	67
F						0	80	75	68	82
G							0	75	77	65
X								0	72	47
Y									0	78
Z										0

The second question relates to the optimal size of this set (of the largest coefficients). As was mentioned taking only a 'small' set enables to develop a metric which is insensitive to small differences but gives more weight to the significant features. On the other hand if the set is too small the metric may not be able to distinguish between images that differ in 'nonsignificant' details. It seems that the answer to this question depends strongly on the application and the data base (in [2] for example, the query process will result in the 20 images that best match the query image rather than a single best match).

The distribution of the largest set of wavelet coefficients at different levels of the wavelet decomposition

The third question relates to the distribution of the indices of the largest coefficients at different levels of the wavelet decomposition. Recalling that at each level of the wavelet decomposition a filter bank composed of a low pass filter and a high pass filter are applied to the data followed by the decimation of the results, the high pass filter at the first level of the decomposition will extract the highest frequency components of the data while the low pass filter will average the data. The second level of the decomposition will remove the next level of high frequency components from the output of the previous low pass filter and so forth, so that for an image with high level of regularity one would expect that the largest wavelet coefficients will appear in the later levels of the decomposition. For an image with random entries one would expect the inverse to happen, i.e. the first levels of the decomposition will contain the largest coefficients.

In an experiment this distribution was calculated for images of size 128x128 using 7 levels of non-standard 'haar' decomposition.

The results are summarized in table 3. They include the standard letters A, B and Z (bitmap images), the standard indexed color images 'woman' and 'wbarb' (from MATLAB) and an indexed color image composed of random entries. The data refers to the distribution of the largest 60 wavelet coefficients.

Table 3 : The distribution of the largest 60 coefficients at different levels of the wavelet decomposition for various images.

	Decomposition Level						
	1	2	3	4	5	6	7
'A'	0	10	15	18	4	12	1
'B'	0	15	20	8	4	12	1
'Z'	0	22	20	5	3	9	1
<i>random_entries</i>	35	22	1	0	1	0	1
'woman'	0	0	5	24	19	8	4
'wbarb'	0	0	12	17	20	7	4

Note that for the image with random entries, almost all the 60 largest coefficients are at the first and second levels of the decomposition. In contrast, for all the other images, none of the coefficients is at the first level of decomposition and for the color indexed images 'woman' and 'wbarb' none is at the first or the second levels of the decomposition. In the standard wavelet decomposition, 3/4 of the total

128^2 coefficients are at level 1 and 15/16 of the coefficients are at levels 1 and 2. This means that for an application with a similar distribution, the search for the largest 60 coefficients can be limited only in the set of coefficients at levels 2 through 7 which is only 1/4 of the total coefficients or even in the set of coefficients at levels 3 through 7 which comprises only 1/16 of the total coefficients, thus increasing the speed of the query algorithm.

The L^q Metric:

In [2], the relative error is based on computing the overlap (or the common elements) of the indices of the 60 largest coefficients in the query and target images. If instead we allow a tolerance on the value of the indices when we compare the common elements, can this improve the metric?. Would this increase the gap between the error in comparing images that closely resemble each other and the error in comparing images which differ significantly? more research needs to be done to see if allowing a tolerance on the value of the indices can improve the metric.

- [1] [1] E. Stollnitz, T. Deroose and D. Salesin, *Wavelets for Computer Graphics*, Morgan Kaufmann Publishers, San Francisco, 1986. pages 43-57.
- [2] [2] C.E. Jacobs, A. Finkelstein and D. Salesin, *Fast Multiresolution Image Querying*, in *Proceedings of SIGGRAPH '95*.
- [3] [3] Rafael C. Gonzalez, *Digital Image Processing*, Addison Wesley, 1993, pp 148-151.
- [4] [4] J. Ritter and G. Wilson, *Handbook of Computer Vision Algorithms in Image Algebra*, CRC Press, 1996, pp 274-276.
- [5] [5] B. Friedman, *Principles and Techniques of Applied Mathematics*, Dover Publications. pages 98-99.
- [6] [6] MATLAB, *Reference Guide*, The Mathworks Inc.
- [7] [7] MATLAB *Wavelet Toolbox*, The Mathworks Inc.
- [8] [8] Gilbert Strang and Truong Nguyen, *Wavelets and Filter Banks*, Wellesley Cambridge. pages 26-27.

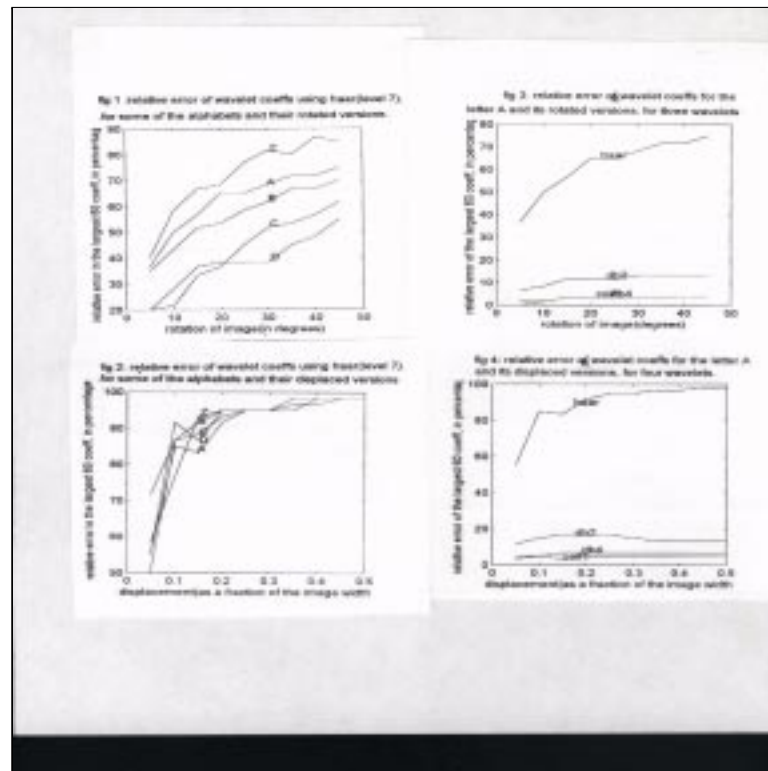


Figure 5.1. Sensitivity of wavelet coefficients to displacement and rotation

Chapter 6

Conclusion

6.1 Summary of the Results

In this study we had employed an innovative method, which has not been used previously in the area of classification of mammographic images. We had extracted feature vectors based on a variant of the approximate Karhunen Loeve Transform adapted to our mammographic images. We had used wavelet packet analysis to derive the Joint best basis, the best possible basis from a library of orthonormal bases to represent a set of images. The Joint best basis and the Karhunen Loeve transform were found to be efficient tools in deriving feature vectors based on the accumulated variance in these bases. We combined our feature vectors with the k-nn as a classifier and achieved results which are better than the average performance of radiologists taking in account that our data base of mammograms came from an early stage of malignancy. Among the various feature vectors and classifiers we used, the performance of the multidimensional K-nn classifier combined with feature vectors based on the accumulated variance in the approximate KLT basis and the Joint best basis (individual bases) were found to provide the best results. We had used the jackknife method to get a robust estimate of the performance of our method. To test the performance of a feature vector, we ran 50 experiments and used the average result to evaluate the feature vector. In each experiment we selected randomly 70% of the data as training data and the rest as test data.

According to studies [34] the average performance of radiologists has a 80% sensitivity and 20% specificity. Compared with these results, our best results have a higher sensitivity (83% compared to 80%) and a higher specificity (55.1% compared to 20%). These results are summarized in the next table:

	sensitivity	specificity
Feature vectors based on approximate KLT	83.2612	55.1209
Average performance of radiologists	80	20

Feature vectors based on the common Joint best bases (common to a set of images, whether all benign, malignant or mixed) did not provide good classification results in our study, though they can serve as good bases for compression and extraction of local features.

The fact that in this study the accumulated variance can serve as a better tool for classification than the variance values themselves is interesting. It seems that the slight differences in the distribution of the variance values have a better contribution to discrimination when they are summed compared to being used individually and separately.

6.2 Classifiers as a second opinion for radiologist

There is extensive work being done in Computer Aided Diagnostics (CAD) in the area of enhancement of mammographic images, automated detection of microcalcifications and masses in mammograms and automated classification of mammographic images into benign and malignant. Having achieved encouraging classification results, our method can provide a second opinion to radiologists in the diagnosis process. The use of the Joint best basis and the approximate Karhunen Loeve transform provide results that are competitive with the results we found in the literature, taking in account that our data base is from an early stage and so is harder for classification.

As a CAD tool for radiologists, the sensitivity of the classifier can be increased (which will result in decreasing the specificity) so that the radiologist can more safely ignore those mammograms classified as benign and examine more closely those mammograms classified as malignant to decide whether to recommend a biopsy or short term follow up examination.

6.3 Future Research

The fact that our method provided better results than the average performance of radiologists is very promising and encourages further research. More research has to be done to improve the robustness of

our results by applying the method to other data bases and using more robust estimation techniques. In a real data set there is a good chance for outliers and therefore robust estimation of the results is important.

As we mentioned, the best basis algorithm selects a basis from a library associated with certain wavelets. Since a mammogram can be analyzed using more than one wavelet function, we can have more than one best basis. Future research may focus on how to optimize the use of more than one best basis as feature vectors.

REFERENCES

- [1] Bird, R.G., T.W.Wallace, and B.C.Yankaskas. *Analysis of cancers missed at screening mammography*. Radiology, 184:613-617, 1992.
- [2] Cox, J.D., S.R.Sharma and R.B. Schilling. *Advanced Digital ammography*. Proceedings of Computer Aided Radiology ' 97, 11-16.
- [3] Duda, Richard O. and Petter H. Hart, *Pattern Classification and Scene Analysis*, John Wiley & Sons, 1973.
- [4] D'Attellis, C.E. and E.M. Fernandez-Berdaguer, Editors, *Wavelet Theory and Harmonic Analysis in Applied Sciences*, Birkhauser, 1997.
- [5] Eckstein, Miguel P. and James S. Whiting, *Lesion Detection in Structured Noise*, Academic Radiology 1995; 2:249-253
- [6] FM, Hall , storella JM, Silverstone DZ, Wyshak G. *Nonpalable breast lessions: recommendation for biopsy based on suspicion of carcinoma at mammography*, Radiology 1988; 167:353-358.
- [7] Friedman, B. , *Principles and Techniques of Applied Mathematics*, Dover Publications.
- [8] Giger, M.L., R.M.Nishikawa, M.Kupinski, U.Bick, et al. *Computerized detection of breast lesions in digitized mammograms and results with a clinically-implemented intelligent workstation*, Proceedings Computer Aided Radiology ' 97 , 325-329.
- [9] Gonzalez, Rafael C. and Richard E. Woods, *Digital Image Processing*, Addison Wesley, 1992.
- [10] Helstrom, Carl W. , *Probability and Stochastic Processes for Engineers*, Macmillan Publishing Company, 1991.
- [11] Horn, Roger A. and Charles R. Johnson, *Matrix Analysis*, Cambridge University Press, 1985.
- [12] Jacobs, C.E. , A. Finkelstein and D. Salesin, *Fast Multiresolution Image Querying*, Proceedings of Special Interest Group for Graphics (SIGGRAPH) '95.
- [13] Jolliffe, I.T., *Principal Component Analysis*, Springer-Verlag, 1986.
- [14] Kaiser, Gerald , *A Friendly Guide to Wavelets*, Birkhauser, 1994.
- [15] Kallergi, M. , L. Clarke, W. Qian, M. Gavrielides, P. Venugopal, C. Berman, S. Holman-Ferries, M. Miller, R. Clark, *Interpretation of Calcifications in Screen/Film, Digitized, and Wavelet-Enhanced Monitor-Displayed Mammograms: A ROC Study*, Academic Radiology 1996; 3:2285-293
- [16] MATLAB, *Reference Guide*, The Mathworks Inc, 1992.
- [17] MATLAB, *Wavelet Toolbox User's Guide*, The Mathworks Inc, 1992.
- [18] MATLAB, *Image Processing Toolbox User's Guide*, The Mathworks Inc, 1992.
- [19] Meisel, William S. , *Computer-Oriented Approaches to Pattern Recognition*, Academic Press, 1972
- [20] Mascio, L., J. Hernandez and C. Logan, Lawrence Livermore National Laboratory, *Automated Analysis for Microcalcifications in High Resolution Digital Mammograms*,
<http://www-sed.llnl.gov/documents/imaging/jmhsie93.html>
- [21] Nishikawa, R.M. , M.L. Giger, et al. *Automated Classification of reast Lesions on Digital Mammograms*. Computer Aided Radiology '97, 347-351.
- [22] Ogden, R. Todd , *Essential Wavelets for Statistical Applications and Data Analysis*, Birkhauser,

1997.

- [23] PGM, Peer *Effect on breast cancer mortality of biennial mammographic screening of women under age 50*, International Journal of Cancer 1995; 60:808:811
- [24] Ritter, Gerhard X. and Joseph N. Wilson, *Handbook of Computer Vision Algorithms in Image Algebra*, CRC Press, 1996.
- [25] Russ, John C. , *The Image Processing Handbook*, CRC Press, 1995
- [26] Saito, Naoki , *Local Feature Extraction and Its Applications Using a Library of Bases*, PhD. Thesis, Yale University, Dec. 1994
- [27] Sayood, Khalid . *Introduction to Data Compression*, Morgan Kaufmann Publishers, Inc. 1996.
- [28] Scharf, Louis L. , *Statistical Signal Processing*, Addison Wesley, 90'.
- [29] Smith, John R. and Shih-Fu Chang, Transform Features for Texture Classification and Discrimination in Large Image Databases, Proceedings of IEEE, Austin, Tx, Nov., 1994
- [30] Stollnitz, E. , T. Deroose and D. Salesin, *Wavelets for Computer Graphics*, Morgan Kaufmann Publishers, San Fransisco, 1986.
- [31] Strang, Gilbert and Truong Nguyen, *Wavelets and Filter Banks*, Wellesley- Cambridge Press, 1996.
- [32] Vidakovic, Brani , Statistical Modeling by Wavelets, *Wiley Series in Probability and Statistics*, 1999.
- [33] *Wavelab Reference Manual*, Stanford University, Ver 0.7, 1995
- [34] Wickerhauser, Mladen V. , *Adapted Wavelet Analysis from Theory to Software*, A K Peters, 1994.
- [35] Wu, Y. C. , M.T. Freedman, A. Hasegawa, R. Zuurbier, Shih-Chung, S.K.Mun, *Classification of Microcalcifications in Radiographs of Pathologic Specimens for the Diagnosis of Breast Cancer*, Academic Radiology 1995; 2:199-204
- [36] Yoshida, H., K. Doi, R. Nishikawa, M. Giger and R. Schmidt. *An improved Computer-Assisted Diagnostic Scheme Using Wavelet Transform for Detecting Clustered Microcalcifications in Digital Mammograms*. Academic Radiology 1996; 3:621-627
- [37] Yoshida, H., R. Nishikawa, M. Giger and K. Doi, *Signal /background separation by wavelet packets for detection of microcalcifications in mammograms*, the International Society for Optical Engineering (SPIE) 96' Vol 2825, pp 805-811.
- [38] Yoshida, H., R. Nishikawa, M. Giger, and K. Doi. Optimally Weighed Wavelet Packets for Detection of Clustered Microcalcifications in Digital Mammograms, DIGITAL MAMMOGRAPHY '96. pp 317-322.
- [39] Kurt Rossmann Laboratories, Department of Radiology, University of Chicago. *Automated Detection of Clustered Microcalcifications*, <http://www-radiology.uchicago.edu/krl/mammo2.1.htm>
- [40] Kurt Rossmann Laboratories, Department of Radiology, University of Chicago. *Automated Detection of Mammographic Masses* <http://www-radiology.uchicago.edu/krl/mammo2.2.htm>
- [41] Kurt Rossmann Laboratories, Department of Radiology, University of Chicago. *Initial Clinical Testing of an 'Intelligent Mammography*, <http://www-radiology.uchicago.edu/krl/mammo2.4.htm>
- [42] Kurt Rossmann Laboratories, Department of Radiology, University of Chicago. *Computerized Classif. of Clustered Microcalcifications*, <http://www-radiology.uchicago.edu/krl/mammo2.6.htm>

BIBLIOGRAPHY

- [1] Bird, R.G., T.W.Wallace, and B.C.Yankaskas. *Analysis of cancers missed at screening mammography*. Radiology, 184:613-617, 1992.
- [2] Cox, J.D., S.R.Sharma and R.B. Schilling. *Advanced Digital ammography*. Proceedings of Computer Aided Radiology ' 97, 11-16.
- [3] Duda, Richard O. and Petter H. Hart, *Pattern Classification and Scene Analysis*, John Wiley & Sons, 1973.
- [4] D'Attellis, C.E. and E.M. Fernandez-Berdaguer, Editors, *Wavelet Theory and Harmonic Analysis in Applied Sciences*, Birkhauser, 1997.
- [5] Eckstein, Miguel P. and James S. Whiting, *Lesion Detection in Structured Noise*, Academic Radiology 1995; 2:249-253
- [6] FM, Hall , storella JM, Silverstone DZ, Wyshak G. *Nonpalable breast lessions: recommendation for biopsy based on suspicion of carcinoma at mammography*, Radiology 1988; 167:353-358.
- [7] Friedman, B. , *Principles and Techniques of Applied Mathematics*, Dover Publications.
- [8] Giger, M.L., R.M.Nishikawa, M.Kupinski, U.Bick, et al. *Computerized detection of breast lesions in digitized mammograms and results with a clinically-implemented intelligent workstation*, Proceedings Computer Aided Radiology ' 97 , 325-329.
- [9] Gonzalez, Rafael C. and Richard E. Woods, *Digital Image Processing*, Addison Wesley, 1992.
- [10] Helstrom, Carl W. , *Probability and Stochastic Processes for Engineers*, Macmillan Publishing Company, 1991.
- [11] Horn, Roger A. and Charles R. Johnson, *Matrix Analysis*, Cambridge University Press, 1985.
- [12] Jacobs, C.E. , A. Finkelstein and D. Salesin, *Fast Multiresolution Image Querying*, Proceedings of Special Interest Group for Graphics (SIGGRAPH) '95.
- [13] Jolliffe, I.T., *Principal Component Analysis*, Springer-Verlag, 1986.
- [14] Kaiser, Gerald , *A Friendly Guide to Wavelets*, Birkhauser, 1994.
- [15] Kallergi, M. , L. Clarke, W. Qian, M. Gavrielides, P. Venugopal, C. Berman, S. Holman-Ferries, M. Miller, R. Clark, *Interpretation of Calcifications in Screen/Film, Digitized, and Wavelet-Enhanced Monitor-Displayed Mammograms: A ROC Study*, Academic Radiology 1996; 3:2285-293
- [16] MATLAB, *Reference Guide*, The Mathworks Inc, 1992.
- [17] MATLAB, *Wavelet Toolbox User's Guide*, The Mathworks Inc, 1992.
- [18] MATLAB, *Image Processing Toolbox User's Guide*, The Mathworks Inc, 1992.
- [19] Meisel, William S. , *Computer-Oriented Approaches to Pattern Recognition*, Academic Press, 1972
- [20] Mascio, L., J. Hernandez and C. Logan, Lawrence Livermore National Laboratory, *Automated Analysis for Microcalcifications in High Resolution Digital Mammograms*,
<http://www-sed.llnl.gov/documents/imaging/jmhsie93.html>
- [21] Nishikawa, R.M. , M.L. Giger, et al. *Automated Classification of reast Lesions on Digital Mammograms*. Computer Aided Radiology '97, 347-351.
- [22] Ogden, R. Todd , *Essential Wavelets for Statistical Applications and Data Analysis*, Birkhauser,

- 1997.
- [23] PGM, Peer *Effect on breast cancer mortality of biennial mammographic screening of women under age 50*, International Journal of Cancer 1995; 60:808:811
 - [24] Ritter, Gerhard X. and Joseph N. Wilson, *Handbook of Computer Vision Algorithms in Image Algebra*, CRC Press, 1996.
 - [25] Russ, John C. , *The Image Processing Handbook*, CRC Press, 1995
 - [26] Saito, Naoki , *Local Feature Extraction and Its Applications Using a Library of Bases*, PhD. Thesis, Yale University, Dec. 1994
 - [27] Sayood, Khalid . *Introduction to Data Compression*, Morgan Kaufmann Publishers, Inc. 1996.
 - [28] Scharf, Louis L. , *Statistical Signal Processing*, Addison Wesley, 90'.
 - [29] Smith, John R. and Shih-Fu Chang, Transform Features for Texture Classification and Discrimination in Large Image Databases, Proceedings of IEEE, Austin, Tx, Nov., 1994
 - [30] Stollnitz, E. , T. Deroose and D. Salesin, *Wavelets for Computer Graphics*, Morgan Kaufmann Publishers, San Fransisco, 1986.
 - [31] Strang, Gilbert and Truong Nguyen, *Wavelets and Filter Banks*, Wellesley- Cambridge Press, 1996.
 - [32] Vidakovic, Brani , Statistical Modeling by Wavelets, *Wiley Series in Probability and Statistics*, 1999.
 - [33] *Wavelab Reference Manual*, Stanford University, Ver 0.7, 1995
 - [34] Wickerhauser, Mladen V. , *Adapted Wavelet Analysis from Theory to Software*, A K Peters, 1994.
 - [35] Wu, Y. C. , M.T. Freedman, A. Hasegawa, R. Zuurbier, Shih-Chung, S.K.Mun, *Classification of Microcalcifications in Radiographs of Pathologic Specimens for the Diagnosis of Breast Cancer*, Academic Radiology 1995; 2:199-204
 - [36] Yoshida, H., K. Doi, R. Nishikawa, M. Giger and R. Schmidt. *An improved Computer-Assisted Diagnostic Scheme Using Wavelet Transform for Detecting Clustered Microcalcifications in Digital Mammograms*. Academic Radiology 1996; 3:621-627
 - [37] Yoshida, H., R. Nishikawa, M. Giger and K. Doi, *Signal /background separation by wavelet packets for detection of microcalcifications in mammograms*, the International Society for Optical Engineering (SPIE) 96' Vol 2825, pp 805-811.
 - [38] Yoshida, H., R. Nishikawa, M. Giger, and K. Doi. Optimally Weighed Wavelet Packets for Detection of Clustered Microcalcifications in Digital Mammograms, DIGITAL MAMMOGRAPHY '96. pp 317-322.
 - [39] Kurt Rossmann Laboratories, Department of Radiology, University of Chicago. *Automated Detection of Clustered Microcalcifications*, <http://www-radiology.uchicago.edu/krl/mammo2.1.htm>
 - [40] Kurt Rossmann Laboratories, Department of Radiology, University of Chicago. *Automated Detection of Mammographic Masses* <http://www-radiology.uchicago.edu/krl/mammo2.2.htm>
 - [41] Kurt Rossmann Laboratories, Department of Radiology, University of Chicago. *Initial Clinical Testing of an 'Intelligent Mammography*, <http://www-radiology.uchicago.edu/krl/mammo2.4.htm>
 - [42] Kurt Rossmann Laboratories, Department of Radiology, University of Chicago. *Computerized Classif. of Clustered Microcalcifications*, <http://www-radiology.uchicago.edu/krl/mammo2.6.htm>