EIGENIMAGE PROCESSING
OF
FRONTAL CHEST RADIOGRAPHS

A thesis

submitted in partial fulfilment

of the requirements for the Degree

of

Doctor of Philosophy

in the

University of Canterbury

by

Anthony P. H. Butler

University of Canterbury

2007

To J.

Abstract

The goal of this research was to improve the speed and accuracy of reporting by clinical radiologists. By applying a technique known as eigenimage processing to chest radiographs, abnormal findings were enhanced and a classification scheme developed. Results confirm that the method is feasible for clinical use.

Eigenimage processing is a popular face recognition routine that has only recently been applied to medical images, but it has not previously been applied to full size radiographs. Chest radiographs were chosen for this research because they are clinically important and are challenging to process due to their large data content. It is hoped that the success with these images will enable future work on other medical images such as those from CT and MRI.

Eigenimage processing is based on a multivariate statistical method which identifies patterns of variance within a training set of images. Specifically it involves the application of a statistical technique called principal components analysis to a training set. For this research, the training set was a collection of 77 normal radiographs. This processing produced a set of basis images, known as eigenimages, that best describe the variance within the training set of normal images. For chest radiographs the basis images may also be referred to as "eigenchests".

Images to be tested were described in terms of eigenimages. This identified patterns of variance likely to be normal. A new image, referred to as the remainder image, was derived by removing patterns of normal variance, thus making abnormal patterns of variance more conspicuous. The remainder image could either be presented to clinicians or used as part of a computer aided diagnosis system.

For the image sets used, the discriminatory power of a classification scheme approached 90%. While the processing of the training set required significant computation time, each test image to be classified or enhanced required only a few seconds to process. Thus the system could be integrated into a clinical radiology department.

Table of Contents

iii

List of Figures

Acknowledgments

I would like to thank my supervisors, Phil Bones and Mike Hurrell, for their guidance and support. Their skills were essential in ensuring that the research remained relevant to both radiology and computer engineering.

I would also like to thank my family, particularly my wife Jenny who has been supportive and encouraging through all the difficult moments. Most importantly she has encouraged me to follow my interests rather than taking easier paths. I would like to thank my sons, Tobin and Dylan, who have had a father who is often busy. I would also like to thank my parents, Pam and Phil, who encouraged me to follow my passion for both computers and radiology.

More widely I would like to thank my friends, especially Struan, Lara, George, and John, who have had to listen to me talk about topics that are often highly technical. I would also like to thank Julian who allowed me to bounce my ideas off him, and who proof read the whole thesis.

I would like to acknowledge the support of various organisations. Firstly, Canterbury District Health Board and The Christchurch Medical School's Department of Radiology who provided access to their PACS archive and other resources, and in particular the special support of Professor Tim Buckenham. Secondly, the Royal Society of New Zealand who have supported this work through their "R.H.C.Bates Award". Similarly, The Royal Australian and New Zealand College of Radiologists have supported the work with their "GE Medical Systems Award" and their "Philips Medical Systems Award". Finally, I would like to thank the Health Research Council of New Zealand who funded me to work full-time on the thesis during 2006.

Chapter I

Introduction

1.1 Aim

The goal of this thesis is to develop and refine a system for computer aided diagnosis of chest radiographs, using eigenimage processing. The system aims to make clinicians faster and more accurate. The work is significant for both its clinical and engineering merits, which are important to the author as a qualified radiologist performing research in the field of computer engineering.

1.2 Clinical significance

The most common examination performed in the majority of radiology departments around the world is the frontal chest radiograph. Previously these have used analogue film and processing, but more recently radiographs have been acquired, processed, stored, and reported using digital techniques and computer-based workstations.

This trend within radiology towards digital images for "plain" radiographs combined with the decreasing cost of computing, has produced a working environment where radiologists can easily access image processing algorithms to help them quickly and accurately report radiographs. Such algorithms are often referred to as Computer Aided Diagnosis (CAD).

Chest radiographs are a very common examination, often performed after hours or at remote locations and the images are then interpreted by non-expert emergency staff or junior doctors from non-radiological specialties. The expert radiologist interpretation is often done hours to days later, often after important medical decisions have been made. Several recent studies comparing emergency specialists to radiologists concluded that there is a significant difference between

radiology specialists and non-specialists in their interpretation of urgent films [1, 2]. Any CAD system that could help non-radiologists with interpretation in the acute setting has the potential to improve patient care.

Even for experts CAD could be of clinical use in improving the accuracy of reporting. For mammography, CAD systems are already in routine use, helping clinicians identify calcifications and masses [3]. Subtraction techniques have been successfully used in chest radiography to aid the reporter in finding lung nodules [4].

1.3 Engineering significance

Recently there has been an increasing use of information theory approaches for pattern recognition algorithms. This trend has been due to the accuracy and flexibility of these approaches. For example, many people have been using training sets of images to identify important patterns, rather than the traditional approach where the investigator chooses which features are important. One successful information theory approach is eigenimage processing. This technique was first used by Turk and Pentland in 1991 for face recognition systems [5, 6].

Engineering stands to benefit from the work in this thesis because it provides insight into the differences and similarities of face recognition problems and CAD systems. The work also provides an example of the application of the technique to unusually large data sets, without taking excessive computational time.

There has been little published work on the application of eigenimage processing to medical images. There are several likely reasons for this:

- Radiographs tend to be very large images. A standard frontal chest radiograph is 2048×2500 pixels, much larger than the images which eigenimage processing was initially applied to. While these radiographs are the largest of radiology's "plain film" images, the volume of data from CT and MRI can be larger still.

- For CAD systems, the recognition problem is different from that of face recognition systems. In CAD systems, there is often one important class of images (the "normal" image) and multiple classes of pathology, with the possibility of each form of pathology being sub-classifiable. Also, unlike

2

face recognition, it is common for a radiograph to belong to several classes at once. For example, a chest radiograph may show both heart failure and pneumonia.

- This use of eigenimage processing also differs from traditional uses in that CAD systems aim to help the clinician make decisions, rather than attempting to replace the clinician. To achieve this a CAD system may simply prioritise images or identify areas of importance on an image, rather than providing a comprehensive classification system.

- Traditionally there has been little overlap between medically trained radiologists and image processing engineers. Fortunately this situation is changing with the development of professional colleges that include both groups.

1.4 Existing CAD systems

Work began on CAD systems soon after the advent of computers. Significant early papers were written in the 1960s [7, 8]. Recently an excellent review of current chest radiograph CAD techniques was published by Ginneken *et al.* [9]. Within clinical practice, mammography has probably received the largest amount of investigation [10, 11].

Existing CAD techniques for chest radiography can loosely be grouped into enhancement, registration, subtraction, region of interest, and segmentation techniques.

Enhancement techniques are methods that allow the data already present to be made more apparent to the viewer. Most available radiological viewing software packages contain several routines for image enhancement, varying from simple contrast enhancement to advanced mathematical methods [12, 13]. One of the best researched techniques is temporal subtraction, which aims to enhance images by removing what was present in previous radiographs [14, 15].

Image registration is the process of aligning two images so that direct comparison is easier. Registering two images and then subtracting them has shown itself to be useful in digital subtraction angiography (DSA) and in the detec-

tion of interval changes, such as the detection of lung nodules and pulmonary opacities.

Region-of-interest analysis and segmentation are techniques for finding features within an image that are useful for the reporting radiologist. This includes lung nodule detection and breast mass/calcification detection [16]. Segmentation can also be used for enhancing structures or removing structures within an image. Methods include thresholding, region growing, edge detection, ridge detection, and fitting of geometric models.

In the last few years there has been a number of large cohort studies that have examined the efficacy of CAD systems in clinical practice. Many of these have demonstrated that CAD systems can successfully improve the identification of nodules and lung opacities [14, 15, 17–19]. While these studies have shown good results, the best success has been in mammography [3, 10, 11, 20–22]. Currently several mammography CAD systems have FDA approval for marketing within the United States of America.

It is worthwhile noting that in mammography the primary job of the radiologist is to find calcifications and masses. These are small objects which have a different signal level from background signal of the image, making feature extraction techniques a viable option. However with frontal chest radiographs, identifying lung nodules is but a small part of the radiologist's job, and identifying common pathologies such as heart failure, pleural effusions, pneumothoraces, and pneumonia requires a more general approach. These pathologies are not subtle in terms of area of the image involved or the signal-to-noise ratio of the abnormality. Rather they often involve changes in the texture or size of anatomical features. It is hoped that recent techniques developed for other pattern recognition systems, such as the eigenimage processing technique, could be a more suitable way to advance CAD systems.

1.5 Eigenimage processing

Eigenimage processing is an information theory approach to image classification. In essence, it is the application of Principal Components Analysis (PCA) to a set of training images to produce a set of basis images that are used for processing of subsequent test images. The method is typically used for classification systems,

4

image compression systems, and image enhancement systems.

In the statistical literature the basis images are known as "principal components", while in the engineering literature they are termed "eigenimages". The word is derived from the fact that the images are eigenvectors of the training set's co-variance matrix. There is a lack of agreement in the literature as to whether the term "eigenimage processing" refers to the application of PCA to any set of images for any purpose, or whether the term should only refer to complete image classification systems in which PCA is a single step.

PCA is a multivariate statistical method for studying correlations between variables [23–25]. It is closely related to the Karhunan-Loève transform and Singular Value Decomposition (SVD) of matrices [26, 27]. The use of PCA in imaging was first proposed by Turk and Pentland in 1991 for face recognition systems [5, 6]. There is also an earlier paper by Kirby and Sirovich studying the use of the Karhunan-Loève transform for characterisation without classification [28].

Eigenimage processing is probably the most frequently used form of image recognition currently in use [29]. Since the initial work of Turk and Pentland on digital pictures of faces, people have built complete face recognition systems with maximum recognition rates close to 90% [30, 31]. It has also been successfully used in optical character recognition, voice recognition, lip-reading, image compression, and image enhancement [32–38].

The technique has shown itself to be most useful for recognising patterns that exist in the whole image, rather than identifying small areas of interest. Given this, it is expected that pathological patterns such as heart failure, pleural effusions, and pneumonias would be more easily recognised than small areas of interest, such as lung nodules.

While its use in medical imaging has been limited, it has been applied to mammography image compression and feature analysis, and regions of interests within CT volume data [39–41]. To date, little work has been published on classification schemes for complete radiological images.

Several variations on eigenimage processing have been described:

- Multi-resolutional approaches where PCA is performed at different resolutions to improve classification or compression [42, 43].

- Non-linear approaches, such as kernel PCA, where the data is transformed

so that the PCA process can find patterns that are non-linear in the original data space [44].

- Iterative approaches, such as incremental PCA and sub-space PCA, where the training set is updated or segmented [29,45]. These approaches allow for the use of large training sets, or alterations of the training set (as required for video compression).

1.5.1 Comparison with other face recognition systems

A good survey of face recognition systems was performed by Chellappa *et al.* [46]. In addition to eigenimage processing (PCA), there are many other statistical methods that have been used. Two other major statistical methods used for face recognition are Independent Component Analysis (ICA) and Linear Discriminant Analysis (LDA). Several authors have attempted direct comparison of these systems using standardised image sets [47–49]. In general, while ICA can be more accurate than PCA for some problems it is often computationally too expensive. For many problems PCA is more accurate than LDA with similar computation required. There have been successful attempts to integrate all three major routines into one system [50].

ICA is a generalisation of PCA. It searches for a linear transformation which expresses the observed data as a linear combination of statistically independent source variables. There are many good papers and tutorials that explain and compare the use of PCA and ICA for face recognition [47, 48, 51]. Due to the computational expense of ICA, some face recognition systems perform PCA for dimensionality reduction prior to applying ICA [51].

In PCA, data is projected into a space where the transformed variables are uncorrelated, that is their co-variance is zero. In ICA, this requirement is strengthened so that the variables are not only uncorrelated, but remain uncorrelated after any transformation of the variables. Sometimes this is summarised by saying that PCA minimises second order statistics (variance), while ICA also minimises higher order statistics [48]. It can be shown that when the observed variables are from a Gaussian distribution, ICA reduces to PCA [51]. Few people have studied the distribution of chest radiographs, but one thorough analysis concluded that a multivariate Gaussian distribution is a reasonable approximation [52]. Thus

for chest radiographs, the computationally more expensive ICA routine is likely to produce the same results as PCA, although no one has formally tested this.

Classification systems based on LDA are often known as either Fisher linear discriminants or Fisher-faces. LDA finds basis vectors that best discriminate between two or more pre-defined classes by considering the scatter between them [53]. Many authors have found that for face recognition, PCA usually outperforms LDA. Others have tried to use non-linear components, such as curvilinear components, for classification but these have failed to become popular, probably due to the computing requirements [54].

In summary, the choice of eigenimage processing (PCA) for chest radiograph analysis is consistent with current trends in face recognition. While both ICA and LDA may offer some benefits, particularly if used in conjunction with PCA, developing a PCA based system first provides a good pathway towards the other forms of analysis.

1.5.2 Method overview

When developing a classification system based on eigenimage processing two sets of images are required. The first group, the training set, is used to calculate the eigenimages. The second group, the test set, is used for evaluating the system.

Turk and Pentland outlined the three basic steps found in most classification systems based on eigenimage processing:

Image registration: All images from both the training set and test image sets are aligned. The aim is to remove differences due to photographic (or in this case radiographic) technique.

Processing of the training set: PCA is applied to the training set to calculate a set of basis images, called eigenimages, that best describe the variance of the training set.

Analysis of test images: Each image from the test set is described as a combination of the eigenimages with one weighting factor per eigenimage. The weighting factors are then used to assign the test image to a particular class. For example, in the case of chest radiographs the weighting factors may be used to assign the test image to be either normal or abnormal.

The three steps are described in more detail below.

1.5.3 Image registration

Image registration is the process of aligning two images with respect to normal structures. In eigenimage processing it is applied to all images in both the training set and test set. The aim is to ensure that any differences between images represent true differences in underlying anatomy and pathology.

Registration is important because no two humans are exactly the same shape and size, nor are two radiological examinations ever performed in a perfectly repeatable manner. When comparing one radiograph with another it is important that any variation observed is the result of pathology, not simply because of differences in the size or position of the patient. Image registration is also important elsewhere in radiological image processing, such as digital subtraction angiography, temporal subtraction, and image segmentation, as well as other types of pattern recognition.

Broadly speaking, image registration systems can be grouped into manual or automatics. For manual registration, a user selects control points that are then aligned. In automatic registration, the computer aligns the images without user intervention. Registration techniques may be simple, and only allow certain transformations such as translation and rotation. Other techniques are more complex and may also allow the images to be scaled or warped.

For the majority of the work in this thesis, registration was performed using an "affine transformation" based on manually selected control points. The images were then cropped, and intensity centred. Intensity centring was performed by removing the empirical mean of the training set from the image.

1.5.4 Processing the training set

This is the core step in eigenimage processing. PCA is applied to a high dimensional data set in order to perform dimensionality reduction. Essentially the method involves writing an $n \times m$ pixel image as a vector of length nm. Possible images then span an nm dimensional space. Actual recorded images are found to be highly correlated, and as such they are clustered in a low dimensional subspace of the entire image space. The aim of PCA is to find an efficient basis for

this sub-space.

The mathematical problem is to find a set of basis vectors (corresponding to basis images) that best represent the variance within the set of training images. Turk and Pentland showed that one such method is to find the eigenvectors of the co-variance matrix. The eigenvectors are ordered by the associated eigenvalues of the co-variance matrix. These eigenvalues are usually referred to as characteristic roots. The eigenvector with the largest characteristic root represents the strongest pattern of variance within the training set.

All principal components exist in the same vector space as the original data, so when the data represents an image, the eigenvectors can also be viewed as images. This has led to the coining of the term "eigenface" by Turk and Pentland [5, 6]. For chest radiographs we have suggested the term "eigenchest". However, since the work presented in this thesis could also be applied to other radiographs, CT images or MRI images, the term "eigenimage" is also used. Figure 1.1 depicts a set of eigenchests formed from a training set of normal chest radiographs.

1.5.5   Analysis of test images

Once the eigenimages have been produced, subsequent test images are described as a linear combination of the eigenimages with associated weighting factors. The weighting factors are then used to classify the test images, by grouping images that are similar into sets. In Turk and Pentland's original paper, classification was performed using the Euclidean length of the weighting factors.

The process of expressing a test image as a linear combination of eigenimages is depicted in figure 1.2. Usually, within each test image there is a small portion of the image that is not able to be represented by the eigenimages. This portion of the test image is referred to as the "null space image". Complementary to this terminology, the portion of the test images described by the eigenimages is often referred to as the "column space image".

Several good reviews describe the types of classification schemes used for eigenimage processing [55, 56]. In particular, they describe the use of "Euclidean distance", "Mahalanobis distance" and "self organising maps". Most methods involve measuring the distance between test images within eigenspace, where each image's location is specified by its weighting factors.
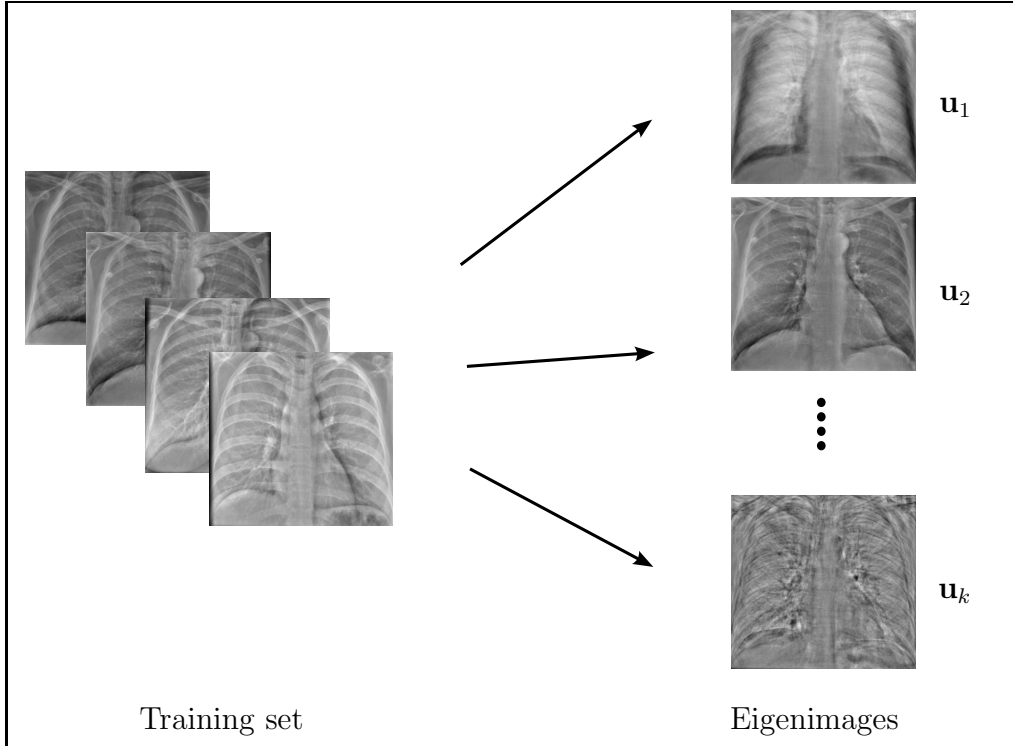
Figure 1.1: **Deriving eigenimages from a training set:** Application of PCA to the training set produces a set of eigenchests, $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k$ that best represent the variance within the set.

A popular variation on eigenimage processing is to define different spaces for each pattern that is to be classified. For example, using the data sets gathered for this thesis, PCA would be performed on both the normal training set and a pneumonia training set. Distance measures, as described above, are then used to test if an image to be classified is closest to the normal set or the pathological set [45, 57, 58]. This approach is similar to LDA.

1.6 Source images

The source images for this thesis were taken from the Christchurch Public Hospital's Picture Archiving and Communication System (PACS). Images were anonymised, then exported in Dicom-3 format, which is an open format supported by most radiological equipment manufacturers [59]. Specifically, images were obtained at a maximum of 2048×2500 pixels at 12 bits per pixel.
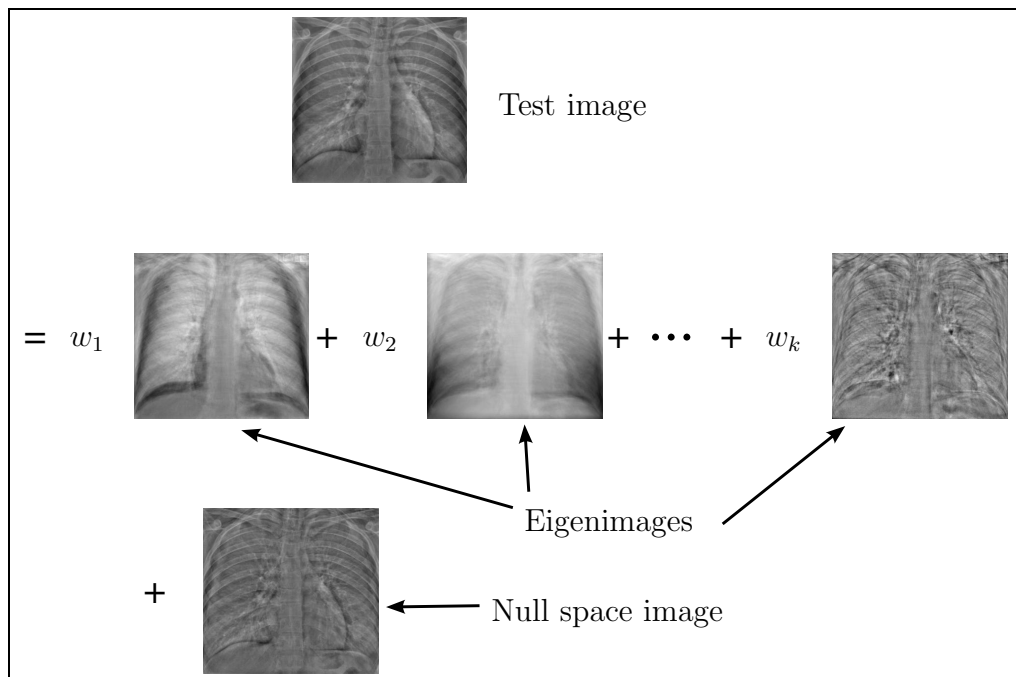
Figure 1.2: **A test image as a linear combination of eigen-images:** Each test image is represented as a linear combination of eigenimages with associated weighting factors. The portion of the test image that can not be represented by the eigenimages is called the null space image.

All images were selected using a key-word search of the radiologist's report stored in the hospital's Radiology Information System (RIS). Two groups of images were obtained: a training set of 77 normal images and a test set of 89 images. The test images were further subclassified into normal or various pathologies (see table 1.1). When initially building the prototype classification system, the test set included only the 15 normal images and 15 images of pneumonia. For later work, the test set was expanded to 89 images by obtaining another 59 images with a variety of other pathologies.

The Canterbury Ethics Committee was informed of the nature of this Ph.D. study in May 2003. The committee noted that all images used were acquired during routine patient management and that they were de-identified by members of the radiology department staff who have access to the images as part of their normal clinical practice. The Committee therefore advised the research group

Training set:		
	77	Normal radiographs
Test images:		
	15	Normal radiographs
	15	Pneumonia
	10	ET tubes, NG Tubes, chest drains and other therapeutic devices
	10	Nodules, carcinoma, and other lung masses
	11	Pneumothoraces of any type
	13	Pleural effusions of any type
	15	Left ventricular failure
Total	89	

Table 1.1: **Diagnosis found on each image:** The groups of chest images used in the study. The pathology represented on each image was taken from the radiologist's report. Training set images were used for construction of eigenimages. Test images are then used for evaluation of the system.

that no ethics approval was necessary for the project.

Before inclusion in the training set all images were manually checked by the author to ensure that all images were high quality PA[1] frontal radiographs. Images were rejected if the subject was overly rotated, poorly inflated, or had the extremities of the chest cavity excluded from the radiograph. Images were also rejected if they contained implanted or overlying devices, such as pacemakers or oxygen masks. ECG[2] electrodes were allowed as these are present on a large fraction of normal radiographs in the hospital.

For the test set, each image was checked by the author to ensure that the image contained only the reported pathology. For the test set both PA and AP[3] images were included. Images were only rejected if part of the subject was excluded from the radiograph. Images containing overlying medical devices

---

[1] Postero-Anterior radiograph: The X-ray beam enters through the patient's back and leaves through their front.

[2] Electrocardiogram: A recording of the electrical activity of the heart.

[3] Antero-Posterior radiograph: The reverse geometry of a PA radiograph.

or which had sub-optimal patient positioning were accepted, as these factors constitute useful information for image interpretation.

As part of the radiographer's normal routine for acquiring chest radiographs, all images are intensity scaled to ensure that they are of a similar brightness, using specific visual criteria. For example, in a properly exposed radiograph the spinal vertebrae should be just visible behind the heart. This scaling of intensities works well because most digital acquisition systems (e.g. CR systems) have a near linear response to X-ray exposure. Apart from the normal radiographer quality assurance, no specific quantitative tests were performed.

1.7   Structure of this thesis

Chapters 2 and 3 review the relevant statistics and algebra. The aim is to give the reader an understanding of PCA for low dimensional problems so they can extrapolate to the high dimensional problems found in eigenimage processing. In particular, chapter 2 aims to familiarise the reader with the concept that principal components (eigenimages) identify the direction of variance, while the associated characteristic roots describe the magnitude of the variance. Chapter 3 highlights the relationship between PCA and the SVD. It emphasises the use of the SVD to solve large PCA problems and discusses how the SVD can provide an algebraic interpretation of PCA.

Chapter 4 describes a simple system for classification of chest radiographs as either normal or abnormal. This prototype system is based on the original face recognition system of Turk and Pentland. The chapter focuses on implementation issues involved with processing very high resolution images. The chapter proves that, with modern computer hardware, eigenimage processing is a feasible technique for chest radiographs.

Chapter 5 focuses on the pre-processing step of aligning all the images. Because medical image registration is a very large topic, the chapter initially outlines a taxonomy of registration systems and reviews commonly used methods. Following this, experiments compare four different semi-automated methods of registration, and assess their quality of alignment. The chapter concludes that each registration method has advantages and disadvantages, with no single method being ideal for all situations.

Chapter 6 addresses the application of PCA to the training set. Experiments study the clustering of training images within the complete image space, examining the dimensionality and spread of the cluster. Experiments then test the effect of training set size on both clustering and computation time. It is inferred that PCA does indeed allow the identification of a cluster of normal images, and allows the cluster to be described to a known level of accuracy, both in dimensionality and spread. The chapter concludes with a discussion of the ideal training set.

Chapter 7 studies the two parts of a test image– the column space image (the portion represented by the eigenimages) and the null space image (the portion not represented by the eigenimages). Initially the chapter describes a classification scheme using Mahalanobis distance to compare column space images. Next, classification is performed using residual analysis to compare the null space images. Both methods are shown to be more accurate for chest radiographs than the prototype system in chapter 4, which is based on Turk and Pentland's face recognition system and used the Euclidean length of the column space image. The chapter also investigates whether displaying the component images is a useful form of image enhancement, to complement the original radiograph by highlighting areas of abnormality, improving the speed and accuracy of the reporting clinician.

The final set of experiments, described in chapter 8, derives a new image that uses information from both the column space and null space components. The new image is referred to as the "remainder image" because it is derived from a test image by subtracting the expected patterns and range of normal variance. Classification based on the remainder image is shown to be better than classification based on the column space or null space images alone. The remainder image is also presented as an enhanced image which could be shown to a reporting clinician. The remainder image is better suited to radiological imaging problems, rather than traditional face recognition problems, as discussed in section 1.1.

1.8 Summary

1. CAD systems aim to improve the speed or accuracy of image interpretation.

2. Eigenimage processing is one of the most successful forms of image pattern recognition. Its use in medical image processing is currently limited, but successful implementation would allow a wide range of applications to be tested. In particular success with chest radiographs would allow the technique to be tested on other radiographs, CT images, and MRI images.

3. This thesis develops an eigenimage processing system for classification and enhancement of frontal chest radiographs. A new derived image is developed which is suited to radiological problems.

4. The goal is to produce a system that will be used in a clinical practice.

Chapter II

Principal components analysis

Principal Components Analysis (PCA) is a form of multivariate statistics. Its goal is to describe a large set of inter-related variables as a much smaller set of uncorrelated variables. It is often referred to as a form of dimensionality reduction because it starts with a set of data with many variables (i.e. many dimensions) and reduces it to a small set of variables (i.e. a smaller number of dimensions). This is achieved using a linear transform. In statistical literature the basis set of vectors for the transformation are called principal components, and they are ordered so that the first few components retain most of the variation present in all of the original variables.

This chapter outlines a derivation of PCA and its most common formulation. It uses a two variable (i.e. two dimensional) numerical example to enable the reader to understand PCA before extrapolating to higher dimensional problems. The chapter initially defines PCA and then introduces the concepts of variance and co-variance as used in multivariate statistics. It then gives two proofs for the theorem that the principal components can be found by calculating the eigenvectors of a sample co-variance matrix. Finally it discusses three important PCA-related topics: spectral decomposition of a co-variance matrix, Mahalanobis distance, and control ellipses.

The majority of the formulation and important results are from an excellent text book by Ingrid Jolliffe [23]. Other good textbooks which cover PCA and multivariate statistics in general are helpful as they highlight different topics and terminology [24, 25, 60, 61]. Jolliffe traces the PCA to the late 19th century when both Beltrami and Jordan independently derived the Singular Value Decomposition (SVD) in a form that underlies PCA [62, 63]. However, it is accepted that the first description of PCA in its modern form was by Pearson in 1901 and later Hotelling in 1933 [64, 65]. As described in the previous chapter, the first use in

imaging was by Turk and Pentland in 1991 who were the first to coin the term eigenimage and eigenface [5, 6]. PCA is also called the Karhunen-Loève transform, after Kari Karhunen and Michel Loève, or the Hotelling transform, after Harold Hotelling.

2.1 Definition of principal components

If there are k random variables then these can be represented using the vector $\mathbf{a}$ of k dimensions to represent this set of variables. The co-variances (or correlations) between the variables are of interest because they describe the relationship between each observed variable. This gives k^2 variances and co-variances to analyse. Therefore if k is very large it will not be helpful to look at these variances alone. Instead, the goal of PCA is to look for a few ($\ll k$) derived variables that preserve as much of the total variance as possible. In other words, the goal of PCA is to find a linear transformation from a high dimensional space to a lower dimensional sub-space while keeping the largest variance of the data.

To achieve this goal look for a linear function which transforms $\mathbf{a}$ onto a new scalar variable $z_1 = \mathbf{u}_1^T \mathbf{a}$, where z_1 has a maximum variance. The vector $\mathbf{u}_1$ is a column of k constants $u_{11}, u_{12}, \ldots, u_{1k}$ giving

$$z_1 = \mathbf{u}_1^T \mathbf{a} = \langle \mathbf{u}_1, \mathbf{a} \rangle = \sum_{i=1}^{k} u_{1i} a_i. \tag{2.1}$$

Next, look for another linear function $\mathbf{u}_2^T \mathbf{a}$ which transforms $\mathbf{a}$ onto the scalar variable z_2 which also has a maximum variance, but with z_2 uncorrelated with z_1. Similarly look for further linear functions $z_i = \mathbf{u}_i^T \mathbf{a}$ which are uncorrelated with the other functions. Up to k derived variables can be found, although for many real world problems most of the variance will be accounted for by fewer derived variables, which is what we are hoping for.

There is some confusion over what to call $\mathbf{u}_i$ and $\mathbf{u}_i^T \mathbf{a}$. Many statisticians call the derived variables z_i the principal components. They refer to $\mathbf{u}_i$ as either the characteristic vector, vector of co-efficients, or vector of loadings, for the ith principal component. In engineering literature $\mathbf{u}_i$ is often directly referred to as the ith principal component, eigenvector, or eigenimage. In Jolliffe's book there is a referenced discussion regarding what is the appropriate terminology [23]. In

this thesis the engineering style nomenclature is used. This is consistent with the style of Turk and Pentland in their pioneering paper on the use of PCA in imaging. Thus $\mathbf{u}_i$ is herein said to be the ith principal component or eigenimage.

An example of 50 observations of two highly correlated random variables is given in figure 2.1. In a real world example this could represent the difference from the population's mean for 50 people's height and weight. This data is then transformed onto a basis which is uncorrelated, plotted in figure 2.2. These transformed variables are $z_1 = \mathbf{u}_1^T \mathbf{a}$ and $z_2 = \mathbf{u}_2^T \mathbf{a}$. In short each observation of $\mathbf{a}$ has been projected onto a new basis $\mathbf{u}_i$ to uncorrelate the observations.

In figure 2.2 there is more variance in the z_1 direction than in the z_2 direction. For larger problems ($k > 2$) there are often significant correlations between the variables and so the first few principal components will account for a majority of the variance. For data where each measured variable represents true signal plus random noise, it is expected that the true signal will be more correlated than random noise, and hence more likely to be represented within low order principal components while the noise is more likely to be represented in the high order principal components.

For an imaging example, consider a set of p images each of $n \times m$ pixels. Thus there are p observations of $k = nm$ random variables, where each of the variables are highly correlated. That is, the brightness of one pixel is correlated with the brightness of other pixels, especially those in its neighbourhood within the image.

2.2   *Variance of multiple random variables*

Before embarking on PCA-related theorems it is important to understand the concepts of co-variance, co-variance matrices, and sample co-variance matrices. Co-variance is a measure of how interdependent two variables are. The co-variance matrix generalises variance from one variable to multiple variables by forming a matrix of co-variances. A sample co-variance matrix is an estimate of the population co-variance matrix found by sampling data.
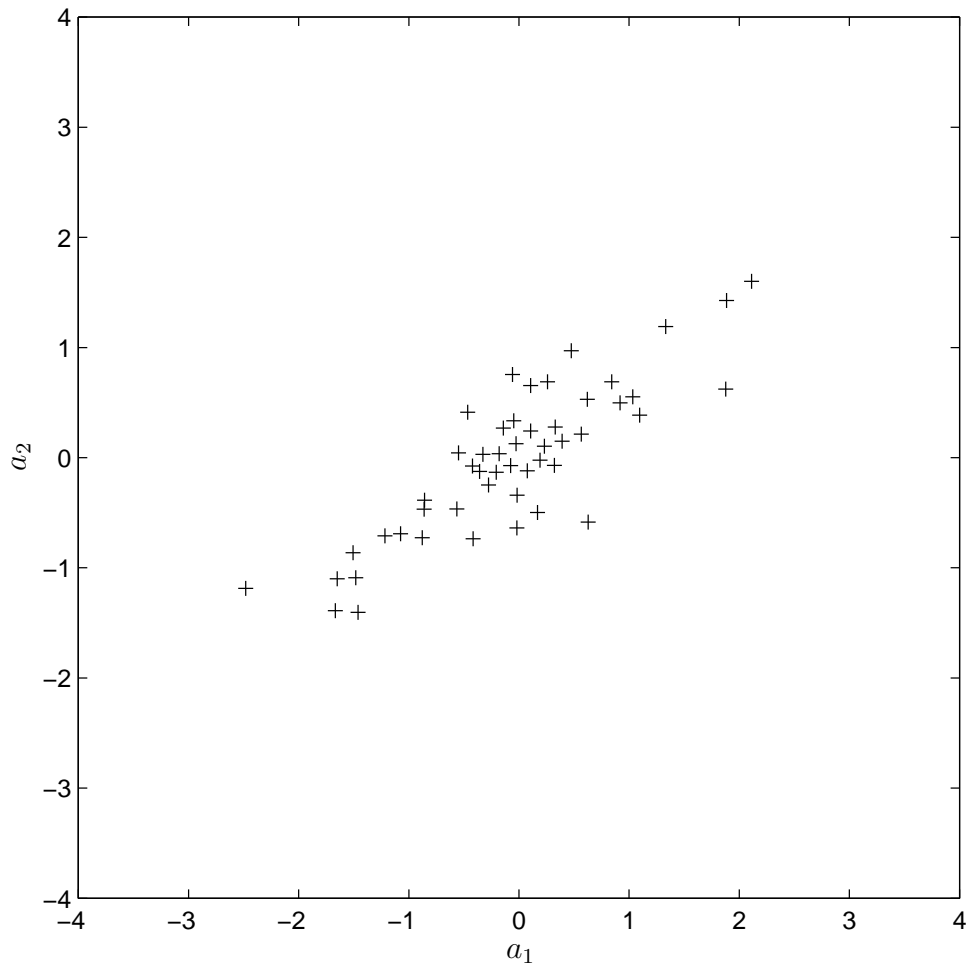
Figure 2.1: **Plot of correlated random variables:** Plot of $p = 50$ observations of **a**. Each observation of **a** has $k = 2$ correlated random variables.
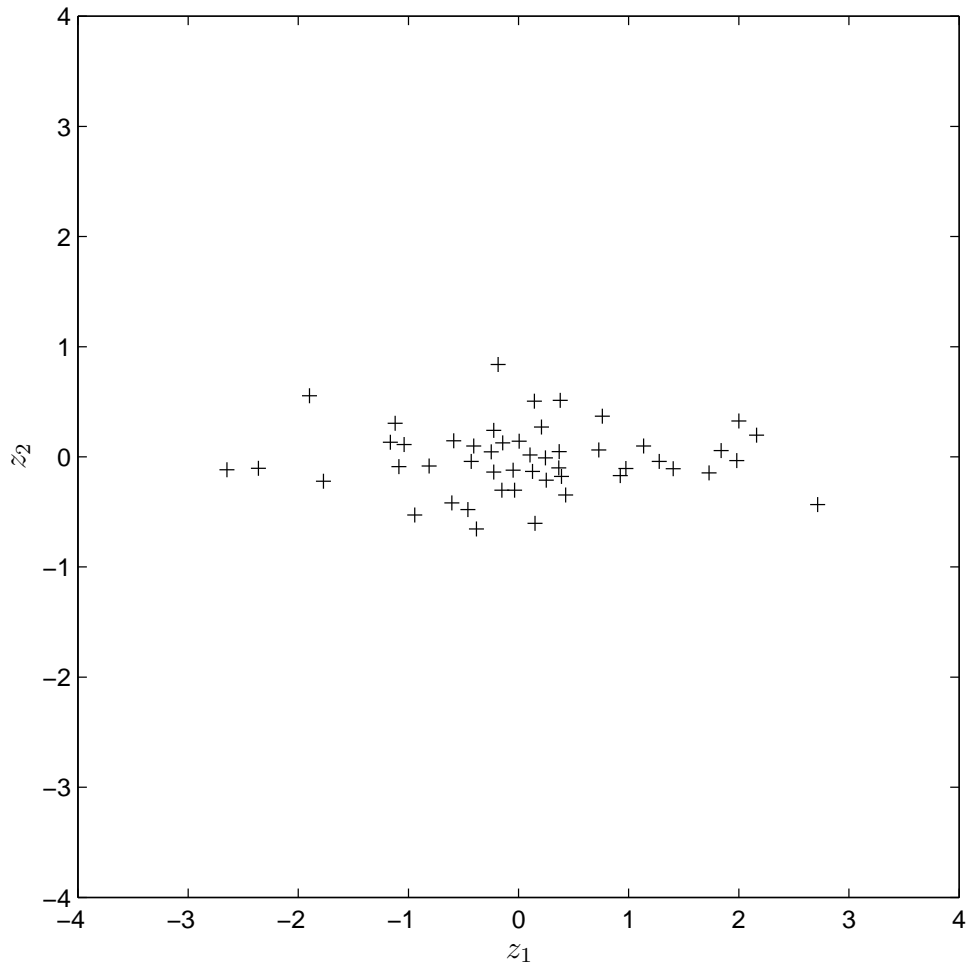
Figure 2.2: **Plot of uncorrelated random variables:** $p = 50$ observations of $k = 2$ correlated random variables after transformation to uncorrelate them. The variables are now $z_1 = \mathbf{u}_1^T \mathbf{a}$ and $z_2 = \mathbf{u}_2^T \mathbf{a}$.

2.2.1 Co-variance

Co-variance indicates the "lack of independence" of the two variables. That is, how two variables vary with each other. A positive co-variance implies that as one variable becomes larger then the other variable also tends to become larger. A negative co-variance implies that as one variable become larger in the positive sense the other variable becomes larger in the negative sense.

Consider the co-variance of two random variables x and y. The co-variance of these variables is formally defined as:

$$\text{cov}\,[x, y] = \text{E}\,[(x - \mu)(y - \nu)] \tag{2.2}$$

where E is the expected value and μ and ν are the expected values (i.e. means) of x and y respectively. This definition can be shown to be equivalent to equation 2.3 for scalar variables and to equation 2.4 for vector variables.

$$\text{cov}\,[x, y] = \text{E}\,[xy] - \mu\nu \tag{2.3}$$

$$\text{cov}\,[\mathbf{x}, \mathbf{y}] = \text{E}\,[\mathbf{x}\mathbf{y}^T] - \mu\nu^T \tag{2.4}$$

For the data in figure 2.1, μ and ν are zero as the data is centred about the origin. This is the reason that in eigenimage processing all images are intensity centred about the mean image.

For most problems the co-variance of two variables is not known. However, it can be estimated by sampling. Specifically for p observations

$$\text{cov}\,[x, y] = \text{E}\,[xy] - \mu\nu \tag{2.5}$$

$$= \frac{1}{p}\sum_{i=1}^{p} x_i y_i - \frac{1}{p^2}\left(\sum_{i=1}^{p} x_i\right)\left(\sum_{i=1}^{p} y_i\right) \tag{2.6}$$

It is worth noting that if x and y are independent then their co-variance is zero. This follows because the definition of independent random variables is:

$$\text{E}\,[xy] = \text{E}\,[x]\,\text{E}\,[y] = \mu\nu \tag{2.7}$$

If x and y are real independent random variables and c is a constant, three other important results should be noted. Equation 2.8 states that the co-variance of a variable with itself is simply the variance of the variable. Equation 2.9 is the commutative property while equation 2.10 is the homogeneity property.

$$\text{cov}\left[x, x\right] = \text{var}\left[x\right] \tag{2.8}$$

$$\text{cov}\left[x, y\right] = \text{cov}\left[y, x\right] \tag{2.9}$$

$$\text{cov}\left[cx, y\right] = c\,\text{cov}\left[x, y\right] \tag{2.10}$$

2.2.2 The co-variance matrix

The co-variance matrix, often denoted Σ, is a matrix of co-variances between elements of a vector. That is, if $\mathbf{a}$ is a column vector of n elements then the co-variance matrix has the $(i, j)^{th}$ element being the co-variance between the ith and j^{th} elements of $\mathbf{a}$ when $i \neq j$ and the variance of the ith element of $\mathbf{a}$ when $i = j$.

Formally the co-variance matrix is defined as:

$$\Sigma = \text{var}\left[\mathbf{a}\right] = \text{E}\left[(\mathbf{a} - \text{E}\left[\mathbf{a}\right])(\mathbf{a} - \text{E}\left[\mathbf{a}\right])^T\right] \tag{2.11}$$

$$= \begin{bmatrix} \text{E}[(a_1 - \mu_1)(a_1 - \mu_1)] & \text{E}[(a_1 - \mu_1)(a_2 - \mu_2)] & \cdots & \text{E}[(a_1 - \mu_1)(a_k - \mu_k)] \\ \text{E}[(a_2 - \mu_2)(a_1 - \mu_1)] & \text{E}[(a_2 - \mu_2)(a_2 - \mu_2)] & \cdots & \text{E}[(a_2 - \mu_2)(a_k - \mu_k)] \\ \vdots & \vdots & \ddots & \vdots \\ \text{E}[(a_k - \mu_k)(a_1 - \mu_1)] & \text{E}[(a_k - \mu_n)(a_2 - \mu_2)] & \cdots & \text{E}[(a_k - \mu_n)(a_k - \mu_k)] \end{bmatrix} \tag{2.12}$$

$$= \begin{bmatrix} \text{cov}[a_1, a_1] & \text{cov}[a_1, a_2] & \cdots & \text{cov}[a_1, a_k] \\ \text{cov}[a_2, a_1] & \text{cov}[a_2, a_2] & \cdots & \text{cov}[a_2, a_k] \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}[a_k, a_1] & \text{cov}[a_k, a_2] & \cdots & \text{cov}[a_k, a_k] \end{bmatrix} \tag{2.13}$$

It is worth noting that this matrix is symmetric and $k \times k$ square. It has variances $\text{cov}[a_i, a_i] = \text{var}[a_i]$ along the diagonal.

Similar to equation 2.4 for co-variance it can be shown that an equivalent definition for the co-variance matrix is:

$$\Sigma = \text{var}[\mathbf{a}] = \text{E}[\mathbf{a}\mathbf{a}^T] - \mu\mu^T \tag{2.14}$$

where μ is the expected value of the $\mathbf{a}$. That is, $\mu = \text{E}[\mathbf{a}] = [\mu_1, \mu_2, \ldots, \mu_k]$

An important result for PCA regarding the co-variance matrix is that for a constant vector $\mathbf{c}$,

$$\text{var}\left[\mathbf{c}^T\mathbf{a}\right] = \mathbf{c}^T \text{var}[\mathbf{a}]\,\mathbf{c}. \tag{2.15}$$

This is similar to equation 2.10.

In this thesis the standard notation $\Sigma = \text{var}[\mathbf{a}]$ is used for the co-variance matrix as it generalises the concept of variance to multiple dimensions and is consistent with the notation $\Sigma = \text{var}[\mathbf{a}] = \text{cov}[\mathbf{a}, \mathbf{a}]$ from equations 2.4 and 2.8. Another common notation is $\Sigma = \text{cov}[\mathbf{a}]$. The latter notation, while confusing, is unambiguous because if "cov" has only one vector argument it refers to the co-variance matrix. Alternatively if it has two scalar arguments then it refers to the co-variance of two variables.

2.2.3 The sample co-variance matrix

In most problems the population co-variance matrix is not known. Normally it must be estimated by sampling real data. This is similar to the sample co-variance given in equation 2.6.

Specifically if there are p samples of $\mathbf{a}$, these can be stored as $\mathbf{a}_i$ where $i = 1, 2, \ldots, p$. Then these samples can be written as a matrix of p column vectors, $A = [\mathbf{a}_1\,\mathbf{a}_2\,\ldots\,\mathbf{a}_p]$. Without loss of generality it can be assumed $\text{E}[\mathbf{a}] = 0$, i.e. the data is centred. Equation 2.4 then gives

$$\Sigma = \text{var}[\mathbf{a}] = \text{E}\left[\mathbf{a}\mathbf{a}^T\right]. \tag{2.16}$$

24

By sampling multiple occurrences of $\mathbf{a}$ this becomes

$$\text{var}[\mathbf{a}] \quad = \quad \frac{1}{p-1} \sum_{i=1}^{p} \mathbf{a}_i \mathbf{a}_i^T \tag{2.17}$$

$$= \quad \frac{1}{p-1} A A^T. \tag{2.18}$$

The denominator term is $p-1$ so that the sample's co-variance is an unbiased estimate of the population's co-variance. For simplicity, in the remainder of this chapter theorems use population co-variance matrices, although corresponding theorems exist for sample co-variance matrices.

2.3   Calculating the principal components

Having defined the principal components and introduced the benefits of finding them, it is now necessary to have a method to calculate them. Firstly the important result is given, then two methods are described which prove the result.

>   **The Result:** For $i = 1, 2, \ldots, k$, the ith transformed variable is given by $z_i = \mathbf{u}_i^T \mathbf{a}$ where the vector $\mathbf{u}_i$, is an eigenvector of the co-variance matrix $\Sigma = \text{var}[\mathbf{a}]$ corresponding to the ith largest eigenvalue λ_i.

Put more simply, the principal components are the eigenvectors of the co-variance matrix. They described a basis for the variables where the variance of each variable is independent. The corresponding eigenvalue, often called the characteristic root, describes the variance for each of the transformed variables.

2.3.1   Justification 1

The following is the method chosen by Jolliffe [23]. It is not a formal proof as it only derives the results for the first two principal components, although higher order principal components can be derived in a similar way. It is given here because it is intuitive and therefore extremely useful in understanding PCA.

First consider the transformed variable $z_1 = \mathbf{u}_1^T \mathbf{a}$. This has variance given by

$$\text{var}\,[z_1] = \text{var}\,\left[\mathbf{u}_1^T \mathbf{a}\right]. \tag{2.19}$$

PCA seeks the vector $\mathbf{u}_1$ which maximises the variance of $\mathbf{u}_1^T\mathbf{a}$. However, this maximisation is not solvable until $\mathbf{u}_1$ is constrained. We choose a normalisation constraint $\mathbf{u}_1^T\mathbf{u}_1 = 1$. That is, the sum of the squares of the elements of $\mathbf{u}_1$ equals 1. Other constraints such as $\max_j [\mathbf{u}_{1j}] = 1$ can be used but are outside the scope of this thesis and readers interested in other constraints are again referred to Jolliffe's text [23].

To perform the maximisation, equation 2.15 is used to express the variance of $\mathbf{u}_1^T\mathbf{a}$ in terms of the co-variance matrix Σ

$$\text{var}\left[\mathbf{u}_1^T\mathbf{a}\right] = \mathbf{u}_1^T\Sigma\mathbf{u}_1 \tag{2.20}$$

To maximise $\mathbf{u}_1^T\Sigma\mathbf{u}_1$ subject to $\mathbf{u}_1^T\mathbf{u}_1 = 1$ the technique of Lagrange multipliers is used. That is, maximise

$$\mathbf{u}_1^T\Sigma\mathbf{u}_1 - \lambda\left(\mathbf{u}_1^T\mathbf{u}_1 - 1\right) \tag{2.21}$$

where λ is the Lagrange multiplier. To find the maximum, look for where the differential with respect to $\mathbf{u}_1$ is zero

$$\Sigma\mathbf{u}_1 - \lambda\mathbf{u}_1 = 0 \tag{2.22}$$

Alternatively

$$\left(\Sigma - \lambda I_k\right)\mathbf{u}_1 = 0 \tag{2.23}$$

where I_k is the $k \times k$ identity matrix. Thus λ is an eigenvalue of Σ and $\mathbf{u}_1$ is the corresponding eigenvector. To find which of the k eigenvectors gives the most variance, return to the quantity to be maximised

$$\mathbf{u}_1^T\Sigma\mathbf{u}_1 = \mathbf{u}_1^T\lambda\mathbf{u}_1 = \lambda\mathbf{u}_1^T\mathbf{u}_1 = \lambda \tag{2.24}$$

This variance will be maximised when λ is the largest eigenvalue. That is, the largest eigenvalue corresponds to the eigenvector which produces the largest variance for $\text{var}\left[\mathbf{u}_1^T\mathbf{a}\right]$. Since λ is the variance of $\mathbf{u}_1^T\mathbf{a}$, it is therefore also the variance of z_1.

For the second principal component, $\mathbf{u}_2$ is the eigenvector that corresponds to the second largest eigenvalue, λ_2, of the co-variance matrix Σ. This is similar

to finding the first principal component except this time maximise $\mathbf{u}_2^T\Sigma\mathbf{u}_2$ with the variance about the component being uncorrelated with the first component. That is when the co-variance between the two is zero, $\text{cov}\left[\mathbf{u}_1^T\mathbf{a}, \mathbf{u}_2^T\mathbf{a}\right] = 0$,

$$\text{cov}\left[\mathbf{u}_1^T\mathbf{a}, \mathbf{u}_2^T\mathbf{a}\right] = \mathbf{u}_1^T\Sigma\mathbf{u}_2 \tag{2.25}$$
$$= \mathbf{u}_2^T\Sigma\mathbf{u}_1 \tag{2.26}$$
$$= \mathbf{u}_2^T\lambda_1\mathbf{u}_1 \tag{2.27}$$
$$= \lambda_1\mathbf{u}_2^T\mathbf{u}_1 \tag{2.28}$$
$$= \lambda_1\mathbf{u}_1^T\mathbf{u}_2. \tag{2.29}$$

Hence

$$\mathbf{u}_1^T\Sigma\mathbf{u}_2 = 0 \quad \text{and} \quad \mathbf{u}_2^T\Sigma\mathbf{u}_1 = 0 \tag{2.30}$$

when $\mathbf{u}_1^T\mathbf{u}_2 = 0$ (that is when the principal components are orthogonal).

Similar proofs exist for $i > 2$. They are not presented here as more general methods exist for finding the ith principal component. Returning to the original diagram of 50 observations of two random variables shown in figure 2.1, and having found $\mathbf{u}_1$, $\mathbf{u}_2$, λ_1, and λ_2, it is possible to add vectors representing $\mathbf{u}_1$ and $\mathbf{u}_2$. This is given in figure 2.3. Since $z_i = \mathbf{u}_i^T\mathbf{a}$ and $\text{var}\left[z_i\right] = \lambda_i$, it can be said that $\mathbf{u}_i$ gives the direction of the variance while λ_i gives the magnitude of the variance.

2.3.2 Justification 2

A second method for establishing that the principal components are eigenvectors of the co-variance matrix is given in this section. While the method establishes the conjecture for all principal components it is less intuitive than the method in section 2.3.1. The best overview of the method is on the world wide web on "Wikipedia" [66]. This method is equivalent to that in Edward Jackson's book "A User's Guide to Principal Components" where he presents a method for diagonalising a co-variance matrix [25].

Consider the matrix version of equation 2.1,

$$\mathbf{z} = U^T\mathbf{a} \tag{2.31}$$
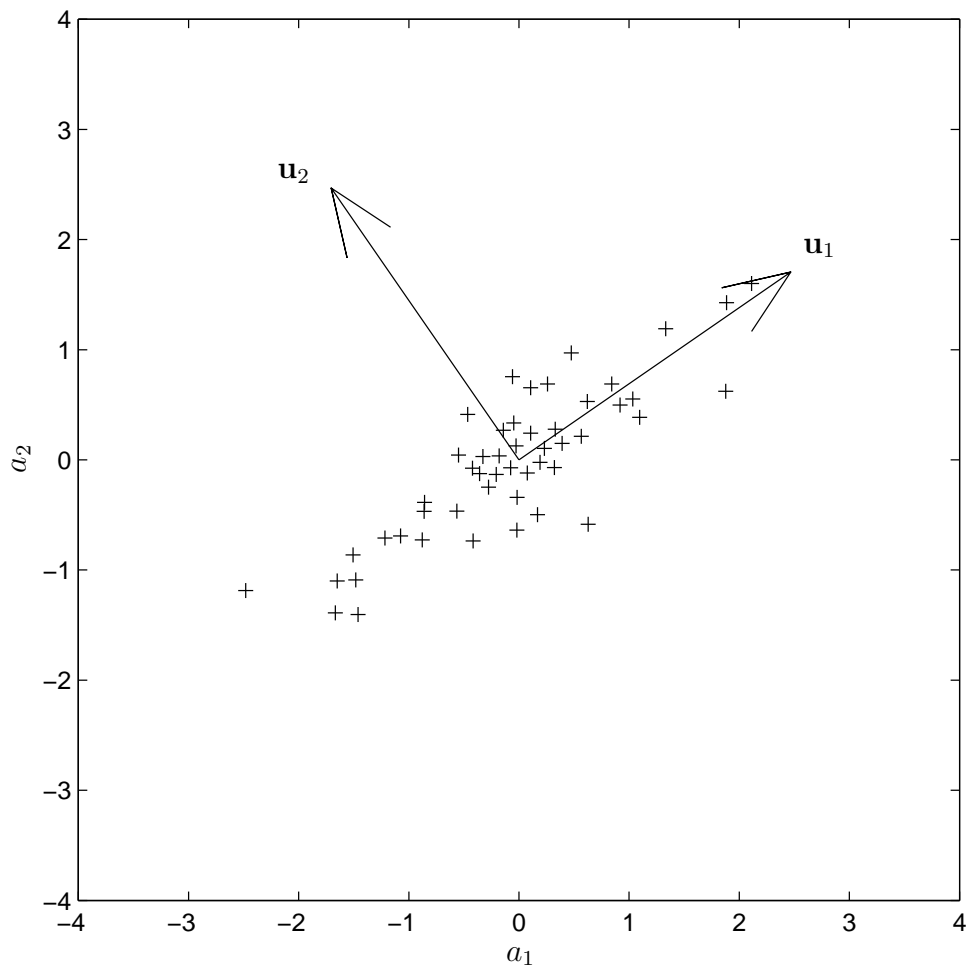
Figure 2.3: **Plot of two correlated random variables with principal components shown:** The plot shows 50 observations of 2 correlated random variables. The principal component vectors have been added to show how $\mathbf{u}_1$ points in the direction of the most variance. Note that $\mathbf{u}_2$ is orthogonal to $\mathbf{u}_1$ and points in a direction that accounts for the remaining variance.

where $\mathbf{a}$ is a vector of k random variables. To simplify the proof, assume $\mathbf{a}$ has a mean of zero, $\mathrm{E}[\mathbf{a}] = 0$. The vector $\mathbf{z}$ represents the transformed variables $z_1 \ldots z_k$. The matrix U is a $k \times k$ projection matrix. We write U as a series of column vectors

$$U = [\mathbf{u}_1 \ \mathbf{u}_2 \ \ldots \ \mathbf{u}_k] \tag{2.32}$$

As discussed previously, the projection of $\mathbf{a}$ (where the variables are correlated) to $\mathbf{z}$ (where they are uncorrelated) is the goal of PCA. The result to be established is that the columns of U are eigenvectors of the co-variance matrix of $\mathbf{a}$. This is done by looking at the co-variance matrix of $\mathbf{z}$, then using substitution and matrix algebra.

Equation 2.31 is constrained using the definition of PCA given in section 2.1: the variables $z_1 \ldots z_k$ must be uncorrelated. This is done by constraining the co-variance matrix $\mathrm{var}[\mathbf{z}] = \Lambda$ to be diagonal. Specifically this ensures that $\mathrm{cov}[z_i, z_j] = 0$ for $i \neq j$ and $\mathrm{cov}[z_i, z_i] = \mathrm{var}[z_i] = \lambda_i$. Thus the co-variance matrix is

$$\Lambda = \begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_k \end{bmatrix}. \tag{2.33}$$

Note that for real values of z_i, $\mathrm{var}[z_i]$ must be positive. Hence $\lambda_i \geq 0$ leading to Λ being positive definite.

U is constrained such that its columns are orthonormal. Thus, U is an orthogonal projection matrix with the following properties

$$\mathbf{u}_i^T \mathbf{u}_j = \begin{cases} 0 \text{ for } i \neq j \\ 1 \text{ for } i = j \end{cases} \tag{2.34}$$

$$I_k = UU^T = U^T U \tag{2.35}$$

Examining Λ, the co-variance matrix of $\mathbf{z}$ gives

$$
\begin{align}
\Lambda &= \text{var}\,[\mathbf{z}] && (2.36) \\
&= \text{E}\,[\mathbf{z}\mathbf{z}^T] && (2.37) \\
&= \text{E}\,\left[(U^T\mathbf{a})\,(U^T\mathbf{a})^T\right] && (2.38) \\
&= \text{E}\,\left[(U^T\mathbf{a})\,(\mathbf{a}^T U)\right] && (2.39) \\
&= U^T\,\text{E}\,[\mathbf{a}\mathbf{a}^T]\,U && (2.40) \\
&= U^T\text{var}\,[\mathbf{a}]\,U && (2.41) \\
&= U^T\Sigma U && (2.42)
\end{align}
$$

Multiplying both sides by U yields

$$ U\Lambda = \Sigma U. \qquad (2.43) $$

Substituting both equations 2.32 and 2.33 into equation 2.43 yields

$$ \left[\; \lambda_1\mathbf{u}_1,\; \lambda_2\mathbf{u}_2,\; \ldots,\; \lambda_k\mathbf{u}_k \;\right] = \left[\; \Sigma\mathbf{u}_1,\; \Sigma\mathbf{u}_2,\; \ldots,\; \Sigma\mathbf{u}_k \;\right]. \qquad (2.44) $$

Comparing like columns gives

$$ \lambda_i\mathbf{u}_i = \Sigma\mathbf{u}_i \quad \text{for } i = 1\ldots k \qquad (2.45) $$

which is the result to be proved, namely that the characteristic roots λ_i and principal components $\mathbf{u}_i$ are the eigenvalues and corresponding eigenvectors of the co-variance matrix $\Sigma = \text{var}[\mathbf{a}]$. If there are only $r < k$ non-zero eigenvalues of Σ then only the first r columns of U correspond to principal components.

2.4   The power spectrum

It is possible to decompose the co-variance matrix Σ into parts constructed of $\mathbf{u}_i$ and λ_i. This technique is known as either "the power spectrum" or "the spectral decomposition" of Σ. It is useful as it allows understanding of how much of the variance and co-variance of $\mathbf{a}$ is represented by each of the principal components.

Reviewing the relationship between the co-variance matrix $\Sigma = \text{var}[\mathbf{a}]$ and

the co-variance matrix for the transformed variables $\Lambda = \text{var}[\mathbf{z}]$ given in 2.43, right multiply both sides by U^T to give

$$\Sigma = U \Lambda U^T \tag{2.46}$$

Given that $\Lambda = \text{var}[\mathbf{z}]$ is a diagonal matrix with λ_i along the diagonal, the right hand side of equation 2.46 can be expanded, giving

$$\Sigma = \sum_{i=1}^{r} \lambda_i \mathbf{u}_i \mathbf{u}_i^T \tag{2.47}$$

This is the spectral decomposition of the co-variance matrix $\Sigma = \text{var}[\mathbf{a}]$. Although the terms are not strictly decreasing, λ_i decreases as i increases, while $\mathbf{u}_i \mathbf{u}_i^T$ remains "about the same size" due to the normalisation constraint. Jolliffe discusses this in detail in Chapter 2 and Section 6.1 of her book [23]. The spectral decomposition demonstrates that studying the eigenvalues λ_i alone determines how much of the variance of $\mathbf{a}$ is accounted for by each principal component [23, 24].

Using all the characteristic roots $\lambda_1 \ldots \lambda_r$, it is possible to calculate the total variance, TV, contained within a co-variance matrix. Total variance is a useful statistic because it can be calculated without computing the entire co-variance matrix. Total variance is usually defined as the sum of the eigenvalues of the co-variance matrix [23].

$$TV(\Sigma) = \sum_{i=1}^{r} \lambda_i \tag{2.48}$$

The right side of the equation is the trace of Σ. The trace of any matrix can be found in two ways, either from the sum of the eigenvalues or the sum of the diagonals [26]. For a co-variance matrix, the diagonal elements are the variance of each variable (see equation 2.13). Hence the total variance can also be found if the variance of the individual variables is known.

$$TV(\Sigma) = \sum_{i=1}^{k} \text{var}[a_i] \tag{2.49}$$

In a similar fashion to total variance, it is possible to determine the amount

of variance represented by only a few low order principal components. The cumulative proportion of total variance (CPTV) spanned by the first q principal components is given by:

$$\text{CPTV}_q = \frac{\sum_{i=1}^{q} \lambda_i}{\sum_{i=1}^{r} \lambda_i} \tag{2.50}$$

It is therefore possible to determine how many principal components to retain in order to span the desired proportion of the total variance. Methods for choosing the number of retained components are often referred to as "stopping rules".

Often it is useful to graphically display the relative importance of the principal components. This can be done using "scree plots", introduced in 1966 by Catell, where the eigenvalues of each component are plotted [67]. Various stopping rules have been described based on the shape of this curve and other parameters. These are discussed further in section 6.3.

2.5 Mahalanobis distance

A common metric for measuring the distance within the space of a vector of random variables is "Mahalanobis distance". The metric attempts to account for any correlations within a multivariate data set and to be scale-invariant for each variable [61, 68, 69]. For an observation $\mathbf{a}_i$ of a vector random variable $\mathbf{a} = [a_1 \dots a_k]$ of mean $\mu = [\mu_1 \dots \mu_k]$ and a co-variance matrix Σ, the Mahalanobis distance is defined as

$$D_M(\mathbf{a}_i) = \sqrt{(\mathbf{a}_i - \mu)^T \Sigma^{-1} (\mathbf{a}_i - \mu)} \tag{2.51}$$

Jackson's text uses the same metric but refers to it as an "overall measure of variability" for use in quality control systems. Both Jackson and Maesschalck *et al.* describe how the measure can be calculated from either the original observation of $\mathbf{a}_i$ or after transformation of the observation into the principal component domain $\mathbf{z}_i$ [25, 69]. The advantage of calculating Mahalanobis distance in the PCA domain is that when Σ is singular the distance can still be calculated. When Σ is singular, it has at least one zero eigenvalue implying Σ^{-1} cannot be found. For eigenimage processing, the sample co-variance matrix is

always singular, because it is formed from p observations of k variables, where $p < k$.

Calculating $D_M(\mathbf{a}_i)$ in the principal components domain requires noting that U is an orthogonal matrix and Λ is diagonal. Equation 2.46 therefore becomes

$$\Sigma^{-1} = U\Lambda^{-1}U^T \tag{2.52}$$

If Σ is singular then Λ will also be singular. In this case the pseudo-inverse of Λ is used for finding Λ^{-1}. This only inverts the first r non-zero elements on the diagonal of Λ.

With centred data (i.e. $\mu = \mathbf{0}$), equation 2.52 can be substituted into equation 2.51 to yield

$$D_M(\mathbf{a}_i) = \sqrt{\mathbf{a}_i^T U \Lambda^{-1} U^T \mathbf{a}_i}. \tag{2.53}$$

Substituting equation 2.31 yields

$$D_M(\mathbf{a}_i) = \sqrt{\mathbf{z}_i^T \Lambda \mathbf{z}_i} \tag{2.54}$$

$$= \sqrt{\sum_{i=1}^{r} \frac{z_i^2}{\lambda_i}}. \tag{2.55}$$

If Σ is non-singular $r = k$, while if it is singular $r < k$.

In summary, Mahalanobis distance is a scalar measure of the distance of a single observation $\mathbf{a}_i$ from the expected value of $\mathbf{a}$. The distance is measured along each of the r (independent) principal components, scaled by the variance along each component. Put another way, Mahalanobis distance transforms the observation from a p-dimensional correlated multivariate distribution to an r-dimensional uncorrelated multivariate distribution with a variance of 1 in each dimension. For this reason, Mahalanobis distance forms the basis of Hotelling's T^2 significance test, where the T^2-distribution is the expected distribution of uncorrelated multivariate data with variance 1 [61, 69].

2.6   Control ellipses

A control ellipse is used to determine a confidence interval around a clustered set of multivariate data. For the example data in figure 2.1 a control ellipse aims to

encompasses a known proportion of the data. Jackson's text is a good reference for control ellipses [25].

Univariate confidence intervals are calculated by the Student t-distribution. A univariate 99% confidence interval is approximately 2.5 standard deviations from the mean, i.e. a width of $2.5\sqrt{\lambda}$. The Student t-distribution requires specification of the number of observations p (degrees of freedom) used to construct the confidence interval. As the sample size approaches infinity the Student t-distribution approximates a normal distribution.

If there are multiple independent variables then the single dimensional Student t-distribution cannot be used to form multidimensional confidence intervals. Instead, a multivariate confidence interval is used. Hotelling's T^2 distribution is a multivariate generalisation of the Student t-distribution and can be used to form multivariate confidence intervals. Its use in multivariate significance tests is similar to the t-distribution's use in univariate significance tests.

To understand why a multivariate form of the t-distribution is used, consider a vector of 100 independent variables. If each of the 100 independent variable uses a one dimensional Student t-distribution to form 100 separate 95% confidence intervals, then it is expected that five variables will lie outside their individual single variable confidence interval. However, if the intention was to form a confidence intervals where 95% of the time all 100 observations fell within the confidence interval, then a different distribution is needed. i.e. rather than finding 100 individual 95% confidence intervals, Hotelling's T^2 distribution aims to find one 95% confidence interval that spans the 100 individual variables 95% of the time.

The T^2 distribution assumes that each variable is independent (i.e. uncorrelated) and of known variance λ_i. Similar to the univariate Student t-distribution, it is both a function of the number of observations p used to construct the confidence interval, and the desired confidence interval α. In addition, T^2 is also a function of the number of independent variables r. Similar to the univariate form, it approaches a multivariate normal distribution as p approaches infinity. It is usually found from either the F-distribution or β-distribution [25, 61, 69, 70].

$$T^2_{r,p,\alpha} = \frac{r(p-1)}{p-r} F_{r,p-r,\alpha} \qquad (2.56)$$

$$\cong \frac{(p-1)^2}{p} \beta_{r/2,(p-r-1)/2,\alpha} \qquad (2.57)$$

The F-distribution and β-distribution are found using lookup tables which are available in most statistical packages. The second form, derived from the β-distribution, is used when the mean and co-variance matrix are unknown, and are therefore found by sampling, which may contain outliers [69]. The confidence interval, or control ellipse, has a length of $\sqrt{T^2\lambda_i}$ on each axis.

To return to the numerical example in figure 2.1, the transformed variables z_1 and z_2 are independent with variances λ_1 and λ_2 respectively. There are $p = 50$ observations with $r = 2$ independent patterns of variance. Choosing an $\alpha = 0.99$ confidence interval, T^2 is found to be 10.36. Thus, the control ellipse for z_1 and z_2 has axes $\sqrt{10.36\lambda_1}$ and $\sqrt{10.36\lambda_2}$ respectively, i.e. the multivariate 99% confidence interval is 3.2 standard deviations from the mean, rather than the 2.5 standard deviations that a univariate distribution would suggest. This confidence interval can be transformed back onto the original variables a_1 and a_2 yielding an ellipse with a major axis along $\mathbf{u_1}$ and minor axis along $\mathbf{u_2}$. This is shown in figure 2.4.

2.7 Summary

1. PCA is a multivariate statistical method in which highly correlated variables are transformed onto a basis of principal components where the variables are uncorrelated.

2. The majority of the variance in the data is in the low order components. This allows high dimensional data to be efficiently represented by a few variables.

3. The principal components can be found by calculating eigenvectors of the co-variance matrix.
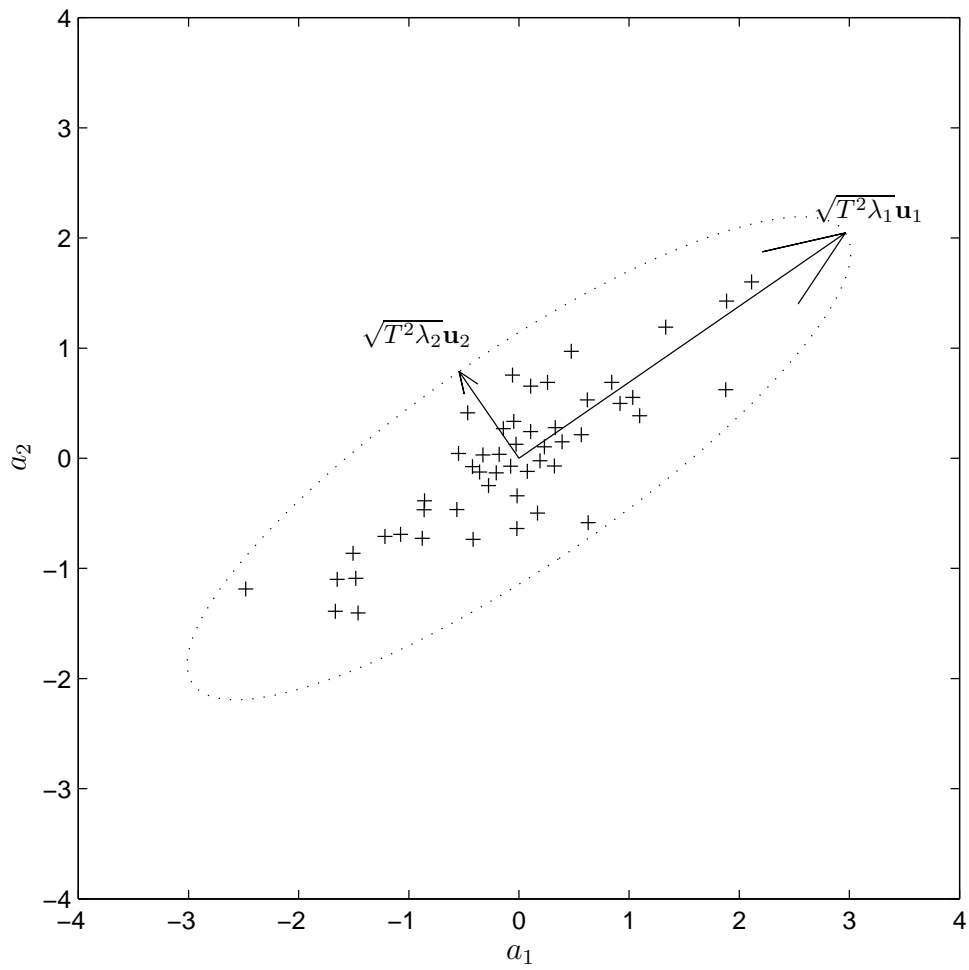
Figure 2.4: **Plot of two correlated random variables with control ellipse:** The plot shows 50 observations of 2 correlated random variables. A 99% control ellipse has been added. This ellipse has axes of length given by $\sqrt{T^2\lambda_i}$

4. The variance associated with each principal component is given by its characteristic root λ_i. This is the eigenvalue corresponding to each eigenvector.

5. PCA can be used to form multivariate confidence intervals known as control ellipses.

Chapter III

The singular value decomposition

The Singular Value Decomposition (SVD) is an algebraic technique for factoring any rectangular matrix into the product of three other matrices. Mathematically and historically, it is closely related to Principal Components Analysis (PCA). It is important because it allows rapid evaluation of the principal components, as required in chapter 2. In addition it provides insight into the geometric interpretation of PCA.

As noted previously, the SVD has long been considered fundamental to the understanding of PCA. In Jolliffe's monograph there is a short section on the SVD [23]. However readers unfamiliar with the SVD are referred to either Strang's text on linear algebra or Golub's text on matrix computations [26, 27]. Within the University of Canterbury there is a good master's thesis covering many important SVD theorems, as well as containing interesting historical information [71]. There are also many good introductions on the world wide web, such as from the University of Wisconsin [72]. The SVD has other imaging applications beyond its use in PCA. These include filtering, watermarking, and compression [73–78].

This chapter begins by defining the SVD, then provides a proof that the SVD can be found for all rectangular matrices. After discussing the algebraic interpretation of the SVD, it introduces the economy size SVD, the singular value spectrum, and effective rank, and finally discusses the relationship between the SVD and PCA.

3.1 Definition of the SVD

The SVD is the factorisation of any $k \times p$ matrix into three matrices each of which have important properties. That is, any rectangular matrix A of k rows

by p columns can be factored in to U, S, and V.

$$A = U\,S\,V^T, \tag{3.1}$$

where

- U is $k \times k$ and has columns which are the eigenvectors of AA^T;

- V is $p \times p$ and has columns which are the eigenvectors of $A^T A$;

- S is a $k \times p$ diagonal matrix with r non-zero singular values on the diagonal, where r is the rank of A. Each singular value is the square root of one of the eigenvalues of both AA^T and $A^T A$. The singular values are ordered so that the largest singular values are at the top left and the smallest singular values are at the bottom right. That, is $s_{1,1} \geq s_{2,2} \geq s_{3,3}$, etc.

Some authors use the terms "left support vectors" and "right support vectors" to refer to the columns of of U and V respectively. A common synonym for singular values is "spread values".

Both U and V have the property of being orthogonal matrices. That is, the columns of U and the columns of V are all orthonormal giving

$$I = \quad UU \quad = U^T U, \tag{3.2}$$
$$I = \quad VV \quad = V^T V, \tag{3.3}$$
$$\therefore S = \quad U^T AV. \tag{3.4}$$

The SVD can be shown to be valid for A having either real or complex elements. In this thesis it is assumed all elements of A are real. Strang gives the corresponding proof for complex A in which transposes are replaced by conjugate transposes [26]. That is A^T is replaced with A^H.

3.2 Proof that all matrices can be factored

This section establishes that any matrix can be factored in to $A = USV^T$ with the properties presented in the previous subsection.

First note that a real symmetric matrix like $A^T A$ has a complete set of orthonormal eigenvectors $\mathbf{v}_i$ which can be stored as the columns of V. Each $\mathbf{v}_i$ is $p \times 1$. Also note that a symmetric matrix has real eigenvalues, γ_i. Both of these results for symmetric matrices are discussed in Strang, section 5.5 [26], and can be summarised:

$$A^T A \mathbf{v}_i = \gamma_i \mathbf{v}_i, \text{ for } i = 1, 2, \ldots, p \tag{3.5}$$

$$\mathbf{v}_i^T \mathbf{v}_j = \begin{cases} 0 \text{ for } i \neq j \\ 1 \text{ for } i = j. \end{cases} \tag{3.6}$$

Taking inner products of equation 3.5 with $\mathbf{v}_i$ demonstrates that $\gamma_i \geq 0$:

$$\mathbf{v}_i^T A^T A \mathbf{v}_i = \gamma_i \mathbf{v}_i^T \mathbf{v}_i \tag{3.7}$$

$$\therefore \|A\mathbf{v}_i\|^2 = \gamma_i. \tag{3.8}$$

Now separate γ_i into the r non-zero eigenvalues $(\gamma_1 \ldots \gamma_r)$, and the remaining $p-r$ values $(\gamma_{r+1} \ldots \gamma_p)$ which are all zero. The latter are solutions to equation 3.7 for which both $A\mathbf{v}_i$ and γ_i are zero.

For the non-zero eigenvalues $(\gamma_1 \ldots \gamma_r)$ set $s_{i,i} = \sqrt{\gamma_i}$ and $\mathbf{u}_i = A\mathbf{v_i}/s_{i,i}$. Note that $\mathbf{u}_i$ are all orthonormal. This is shown by expanding the inner product, then substituting equation 3.7,

$$\mathbf{u}_i^T \mathbf{u}_i = \frac{(A\mathbf{v}_i)^T A\mathbf{v}_i}{s_{i,i}^2} \tag{3.9}$$

$$= \frac{\mathbf{v}_i^T A^T A\mathbf{v}_i}{s_{i,i}^2} \tag{3.10}$$

$$= \frac{\gamma_i \mathbf{v}_i^T \mathbf{v}_i}{s_{i,i}^2} \tag{3.11}$$

$$= \mathbf{v}_i^T \mathbf{v}_i \tag{3.12}$$

$$= \begin{cases} 0 \text{ for } i \neq j \\ 1 \text{ for } i = j. \end{cases} \tag{3.13}$$

The r orthonormal basis vectors $(\mathbf{u}_1 \ldots \mathbf{u}_r)$ can be extended by Gram-Schmidt orthogonalisation to a complete orthonormal basis of k vectors. This is covered in

chapter 3 of Strang's text [26]. In brief, the Gram-Schmidt orthogonalisation is performed by first choosing a vector $\mathbf{b}$, then subtracting from $\mathbf{b}$ all components in the directions of $\mathbf{u}_1 \ldots \mathbf{u}_r$. Thus, $\mathbf{b}$ becomes orthogonal to all the vectors $\mathbf{u}_1 \ldots \mathbf{u}_r$. Then the vector $\mathbf{b}$ is normalised and stored as $\mathbf{u}_{r+1}$. The process is repeated to find $\mathbf{u}_{r+2}$ to $\mathbf{u}_k$. Thus, the vectors $\mathbf{u}_1 \ldots \mathbf{u}_k$ will all be orthonormal and can form columns of U.

It is now possible to find of all the elements of S. Starting with an element-wise form of equation 3.4, then examining the results for $j > r$ and $j \leq r$ gives

$$s_{i,j} = \mathbf{u}_i^T A \mathbf{v}_j. \tag{3.14}$$

$$= \begin{cases} \mathbf{u}_i^T s_{j,j} \mathbf{u}_j & \text{for } j \leq r \text{ (because } A\mathbf{v}_j = s_{j,j}\mathbf{u}_j) \\ 0 & \text{for } j > r \text{ (because } A\mathbf{v}_j = 0) \end{cases} \tag{3.15}$$

The columns of U are orthonormal, implying $\mathbf{u}_i^T \mathbf{u}_j = 0$ except when $i = j$. Thus for $S = U^T A V$ the only non-zero elements are the first r elements on the diagonal, i.e. when $i = j$ and $j \leq r$. These non-zero elements are $s_{i,i} = \sqrt{\gamma_i}$ for $i = 1 \ldots r$. Since S is a diagonal matrix, for convenience the singular values $s_{i,i}$ are often simply written as s_i.

Any rectangular matrix A of k rows and p columns can now be written as

$$A = U S V^T \tag{3.16}$$

where

$$A = \begin{bmatrix} a_{11} & \ldots & a_{1p} \\ \vdots & & \vdots \\ a_{k1} & \ldots & a_{kp} \end{bmatrix} \tag{3.17}$$

$$U = [\mathbf{u}_1 \ldots \mathbf{u}_r \, \mathbf{u}_{r+1} \ldots \mathbf{u}_k] \tag{3.18}$$

$$S = \begin{bmatrix} s_i & & & \\ & \ddots & & \\ & & s_r & \\ & & & \ddots \end{bmatrix} \tag{3.19}$$

$$V = [\mathbf{v}_1 \ldots \mathbf{v}_r \, \mathbf{v}_{r+1} \ldots \mathbf{v}_p]. \tag{3.20}$$

3.3 Algebraic interpretation

The SVD is useful because it gives important information about the $k \times p$ matrix A. In particular, it gives an orthonormal basis for four fundamental spaces of A. Therefore it also reveals the rank of A. Furthermore it provides a way to view A as a map between the row space and column space of A.

3.3.1 The column space

The column space of a matrix is the sub-space spanned by the column vectors of the matrix. Formally the column space of A is all values of $\mathbf{y}$ for which $A\mathbf{x} = \mathbf{y}$ can be solved. It is known that all solutions to $A\mathbf{x}$ are combinations of the columns of A. It is also known that the first r columns of U provide an orthonormal basis for the columns of A. Thus, the SVD provides the first r columns of U which form the orthonormal basis for the column space of A.

If A has k rows, then the column space is an r-dimensional sub-space within the k-dimensional space of real numbers, i.e. $\mathbf{R}^k$.

3.3.2 The left null space

The left null space is the remaining sub-space of $\mathbf{R}^k$ not within the column space. With reference to the SVD it is the remaining $k - r$ columns of U, $\mathbf{u}_{k-r} \ldots \mathbf{u}_k$. Formally it is defined as the solutions for $\mathbf{z}$ such that $\mathbf{z}^T A = 0$.

3.3.3 The row space

The row space of A is the column space of A^T. It is the space spanned by the rows of A. It is an r-dimensional sub-space in $\mathbf{R}^p$. Similar to the column space the SVD gives an orthonormal basis for the row space, namely $\mathbf{v}_1 \ldots \mathbf{v}_r$.

3.3.4 The null space

The null space of A is the left null space of A^T. It is the sub-space in $\mathbf{R}^p$ not in the row space. The SVD provides an orthonormal basis for the null space, namely $\mathbf{v}_{p-r} \ldots \mathbf{v}_p$.

3.3.5 Rank

The rank of a matrix is defined as the maximal number of columns (or rows) of A which are linearly independent. That is, it is the dimensionality of both the row space and column space. Clearly from the SVD this is r, the number of non-zero eigenvalues of AA^T and $A^T A$.

3.3.6 The economy size SVD

When considering the SVD of matrices for which the null space and left null space are non-zero, it is often convenient to ignore the columns of U and V that correspond to these null spaces. Several names have been given to this form of the SVD; MATLAB refers to it as the "economy size" SVD while others refer to it as the "thin" SVD [27, 79, 80].

Consider the SVD of A

$$A = [\mathbf{u}_1 \dots \mathbf{u}_r \, \mathbf{u}_{k-r} \dots \mathbf{u}_k] \begin{bmatrix} s_i & & & \\ & \ddots & & \\ & & s_r & \\ & & & \ddots \end{bmatrix} [\mathbf{v}_1 \dots \mathbf{v}_r \, \mathbf{v}_{p-r} \dots \mathbf{v}_p]^T. \quad (3.21)$$

It is possible to ignore the columns of U and V which correspond to zero values on the diagonal of S. This leads to

$$A = [\mathbf{u}_1 \dots \mathbf{u}_r] \begin{bmatrix} s_i & & \\ & \ddots & \\ & & s_r \end{bmatrix} [\mathbf{v}_1 \dots \mathbf{v}_r]^T \quad (3.22)$$

$$A = U_r \, S_r \, V_r \quad (3.23)$$

where U_r is a matrix formed from the first r columns of U (corresponding to the column space), V_r is a matrix formed from the first r columns of V (corresponding to the row space), and S_r is the square diagonal matrix with the non-zero singular values on the diagonal. Note that U_r is of k rows and r columns, V_r is of p rows and r columns, and S_r is square with r rows and r columns.

3.3.7 Linear transforms

Every linear transformation from $\mathbf{R}^p$ to $\mathbf{R}^k$ can be represented by a matrix A of k rows and p columns. Such transformations are often referred to as linear maps or linear operations. A linear transform is one that preserves addition and scalar multiplication. This may be written as $A\mathbf{x} = \mathbf{y}$, where $\mathbf{x} \in \mathbf{R}^p$ and $\mathbf{y} \in \mathbf{R}^k$.

Any linear transform (e.g. A) maps the vector's components which are in the row space of A, onto the column space of A. However the vector's components within the null space of A are mapped onto zero.

If the SVD is used to factorise A, the operation $A\mathbf{x}$ can be viewed as comprising a projection of $\mathbf{x}$ onto the orthonormal basis of the row space $(\mathbf{v}_1 \ldots \mathbf{v}_r)$, then mapping the resulting vector onto the orthonormal basis of the column space $(\mathbf{u}_1 \ldots \mathbf{u}_r)$ with s_i as a scale factor. This can be shown formally by considering the transformation $A\mathbf{x} = \mathbf{y}$ using the economy sized SVD. The transform is then

$$\mathbf{y} = A\mathbf{x} \tag{3.24}$$

$$= U S V^T \mathbf{x} \tag{3.25}$$

$$= [\mathbf{u}_1 \ldots \mathbf{u}_k]
\begin{bmatrix}
s_i & & & \\
& \ddots & & \\
& & s_r & \\
& & & \ddots
\end{bmatrix}
\begin{bmatrix}
\mathbf{v}_1^T \mathbf{x} \\
\vdots \\
\mathbf{v}_p^T \mathbf{x}
\end{bmatrix} \tag{3.26}$$

$$= \sum_{i=1}^{r} s_i \left(\mathbf{v}_i' \mathbf{x} \right) \mathbf{u}_i. \tag{3.27}$$

This result makes it clear why the SVD is so important, as it establishes which components of $\mathbf{x}$ are in the row space. These components are the only part which are mapped to the column space, whereas the components of $\mathbf{x}$ in the null space correspond to zero values of s_i and are therefore mapped onto zero.

It is worth noting that the pseudo-inverse of A is based on the reverse mapping of the column space back to the row space, with the left null space mapped to zero. The pseudo-inverse is the best "least squares" inversion of a singular matrix. It is normally constructed using the SVD and is described in detail in Appendix 1 of Strang's text [26].

3.4 The singular value spectrum

As shown in section 2.4, it is possible to represent any symmetric matrix, such as a co-variance matrix, as a spectral decomposition. A similar spectrum can be found for non-symmetric matrices such as A. This spectrum is formed using the SVD [27, 81]. Equation 3.22 gives

$$A = \sum_{i=1}^{r} s_i \mathbf{u}_i \mathbf{v}_i^T. \tag{3.28}$$

Since $\mathbf{u}_i$ and $\mathbf{v}_i$ are of unit length, $\mathbf{u}_i \mathbf{v}_i^T$ remains "approximately the same size" for all i. It is possible to ignore contributions to A from terms with $i > r$, since $s_i = 0$ for $i > r$.

Using the singular value spectrum it is possible to gauge the relative importance of components in determining the matrix A. Looking only at the singular values (i.e. the square root of the eigenvalues) it is possible to determine the relative importance of each eigenvector in determining the full matrix A.

The singular value spectrum of A is very closely related to the power spectrum of the co-variance matrix AA^T. This can be shown by right multiplying both sides of equation 3.28 by A^T.

$$
\begin{aligned}
AA^T &= \left(\sum_{i=1}^{r} s_i \mathbf{u}_i \mathbf{v}_i^T \right) \left(\sum_{i=1}^{r} s_i \mathbf{u}_i \mathbf{v}_i^T \right)^T &(3.29) \\
&= \left(\sum_{i=1}^{r} s_i \mathbf{u}_i \mathbf{v}_i^T \right) \left(\sum_{i=1}^{r} s_i \mathbf{v}_i \mathbf{u}_i^T \right) &(3.30) \\
&= \sum_{i=1}^{r} s_i^2 \mathbf{u}_i \mathbf{u}_i^T \quad \text{(from orthogonality of } \mathbf{v}_i \text{ in eqn. 3.6)} &(3.31) \\
&= \sum_{i=1}^{r} \gamma \mathbf{u}_i \mathbf{u}_i^T &(3.32)
\end{aligned}
$$

which is the power spectrum described in section 2.4.

3.5 Effective rank

As previously discussed, the rank of a matrix is the dimensionality of the row space (or column space). Formally, it is the number of independent rows (or columns) of a matrix. For many matrices where the elements are measured experimentally this definition is too simplistic owing to the presence of noise or rounding errors in calculations. Consider the matrix

$$A = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}. \tag{3.33}$$

This matrix has a rank of 1. As expected, it has two eigenvalues, namely 5 and 0 leading to one non-zero singular value, again implying a rank of $r = 1$.

Now consider the same matrix with noise added,

$$A_n = \begin{bmatrix} 1.1 & 2.1 \\ 1.9 & 4.1 \end{bmatrix}. \tag{3.34}$$

Formally this has a rank of 2. Its eigenvalues are approximately 5.1 and 0.1, leading to $r = 2$ non-zero singular values. However it would be preferable to acknowledge that the second non-zero eigenvalue is due to noise in the data, and that the rank of A_n should be 1.

The goal of the effective rank is decide at what value one should consider small singular values to be equivalent to zero for the purpose of calculating the rank of A. In the example using A_n, the matrix with noise added, it may be useful to choose a tolerance where eigenvalues less than 0.2 are considered zero, thus allowing for noise in the system. Such a tolerance would make the effective rank one. Effective rank is usually denoted by ϵ.

In many cases the appropriate method for choosing the tolerance for zero singular values is not clear. A method for choosing the tolerance in PCA is introduced in section 3.6, then evaluated experimentally later in chapter 6.

Once the effective rank of A is chosen it is possible to write the economy size SVD with a rank of ϵ

$$A \approx A_\epsilon = U_\epsilon S_\epsilon V_\epsilon \tag{3.35}$$

where $U_\epsilon = [\mathbf{u}_1 \dots \mathbf{u}_\epsilon]$, $V_\epsilon = [\mathbf{v}_1 \dots \mathbf{v}_\epsilon]$, and S_ϵ is a diagonal matrix with the

elements $s_1 \ldots s_\epsilon$.

A_ϵ is the closest rank-ϵ matrix to A (in a least squares sense). This can be seen by considering the spectral form of the SVD

$$A_\epsilon = \sum_{i=1}^{\epsilon} s_i \mathbf{u}_i \mathbf{v}_i^T. \tag{3.36}$$

Increasing ϵ progressively adds smaller and smaller amounts to A. Readers requiring a formal proof that this is indeed the best least squares approximation are referred to Golub's text [27].

3.6 Use of the SVD in PCA

As stated earlier, the SVD has long been known to be closely related to PCA. Jolliffe gives an account of the historical derivations for the SVD and PCA, while other authors have discussed the relationship in more detail [23, 25, 81–83].

The SVD is helpful in performing PCA when the co-variance matrix is very large but of low rank, i.e. when $p \ll k$. In addition, PCA theorems for determining the number of significant principal components provide a method for determining the effective rank of A.

The relationship can be observed by returning to PCA where the goal is to find the eigenvectors (principal components) and eigenvalues (characteristic roots) of a sample co-variance matrix. Thus PCA attempts to find $\mathbf{u}_i$ and λ_i from

$$\lambda_i \mathbf{u}_i \;=\; \Sigma \mathbf{u}_i \tag{3.37}$$

$$\;=\; \frac{1}{p-1} A A^T \mathbf{u}_i \tag{3.38}$$

for the $i = 1 \ldots r$ non-zero eigenvalues.

Similarly, one of the results from the SVD is the relationship between the column space $(\mathbf{u}_1 \ldots \mathbf{u}_r)$, and singular values $(s_i = \sqrt{\gamma_i})$ of A. Namely,

$$s_i^2 \mathbf{u}_i = A A^T \mathbf{u}_i \tag{3.39}$$

for the $i = 1 \ldots r$ non-zero singular values.

3.6.1 Principal components analysis with large values for k

In chapter 2, a random variable $\mathbf{a}$ of k elements was defined. Taking p observations of $\mathbf{a}$ allowed formation of a data matrix in which each observation was stored as a column of A. Thus, $A = [\mathbf{a}_1 \ldots \mathbf{a}_p]$ where A had p columns and k rows. For the eigenimaging process, introduced in chapter 1, typical values are $p = 100$ training images, each of $k = 3 \times 10^6$ pixels per image. Thus $p \ll k$.

When performing PCA the sample co-variance matrix $\Sigma = \frac{1}{p-1} A A^T$ is examined. This matrix has $k \times k$ dimensions, so therefore, if k is large then $A A^T$ cannot be directly calculated or stored on a computer, and nor can its eigenvalues be found. However the matrix $A^T A$ has $p \times p$ elements. Hence, if p is small, $A^T A$ can be calculated allowing V and S to be found. Using V and S the first r columns of U can then be calculated. That is,

$$\mathbf{u}_i = A \mathbf{v}_i s_i^{-1} \quad \text{for } i = 1 \ldots r. \tag{3.40}$$

Hence using the SVD, the principal components can all be calculated without the need to calculate the entire sample co-variance matrix.

To calculate the characteristic roots, note that the singular values in the SVD are the square roots of the non-zero eigenvalues of $A A^T$, while the characteristic root of the sample co-variance matrix are the non-zero eigenvalues of $\Sigma = \frac{1}{p} A A^T$. From equations 3.38 and 3.39 it is possible to derive a relationship between the singular values of A and the characteristic roots of Σ

$$\lambda_i = \frac{s_i^2}{p-1}. \tag{3.41}$$

Interestingly, once the complete set of $\mathbf{u}_1 \ldots \mathbf{u}_r$ has been found the complete economy sized SVD will have been calculated ($A = U_r S_r V_r^T$). This highlights that the principal components of $\mathbf{a}$ can be considered an efficient orthonormal basis for the observations of $\mathbf{a}$.

3.6.2 Effective rank in the context of PCA

In PCA it is important to determine the effective rank of the sample co-variance matrix Σ. The effective rank of Σ is also the effective rank of A. The SVD

allows for the determination of the effective rank of A by examining the number of significant singular values. From a PCA perspective, the effective rank determines the number of principal components that represent true patterns of variance, rather than components that arise due to noise in the data. From an SVD perspective, the effective rank determines how much of the data matrix is real signal, versus how much is noise.

As discussed in section 2.4, examining the eigenvalues associated with each principal component determines of how much of the total variance is represented by a given number of principal components. For example, choosing enough components to represent 95% of the variance may be considered the appropriate tolerance for the problem. The number of eigenvalues required to cover 95% of the variance is then the effective rank, denoted ϵ, of A. It is usually less than the true rank, r.

In a real world problem with p observations of k random variables it is likely (owing to noise) that there are $r = \min(p, k)$ non-zero singular values. However, using a stopping rule, such as the power spectrum discussed in section 2.4, it may be determined that $\epsilon < r$ is an appropriate effective rank to cover 95% of the variance in the set. In this example $\mathbf{u}_1 \ldots \mathbf{u}_\epsilon$ would represent true principal components while $\mathbf{u}_{\epsilon+1} \ldots \mathbf{u}_r$ represent noise. This is the method often used for SVD based image compression and noise filtering [73–75, 77, 78].

3.7 Notation for this thesis

Later in this thesis it is found that the column space and left null space are the most useful of the four fundamental sub-spaces. In addition, it is found that the effective rank of A is more important the actual rank. Hence the following notation is used for these important sub-spaces:

$\mathcal{C}(A)$: The column space of A. Its dimensionality is the effective rank ϵ.

$\mathcal{N}(A)$: The left null space of A, with dimensionality $k - \epsilon$.

The two spaces can be defined by their basis vectors

$$\mathcal{C}(A) = \{\mathbf{u}_1, \, \mathbf{u}_2, \, \ldots, \, \mathbf{u}_\epsilon\}, \tag{3.42}$$

$$\mathcal{N}(A) = \{\mathbf{u}_{\epsilon+1}, \, \mathbf{u}_{\epsilon+2}, \, \ldots, \, \mathbf{u}_k\}. \tag{3.43}$$

The addition of the column space and left null spaces forms the complete k-dimensional space of real numbers, $\mathbf{R}^k$. Also, the column space is orthogonal to the left null space. These two relationships are represented as

$$\mathbf{R}^k = \mathcal{C}(A) \oplus \mathcal{N}(A) \tag{3.44}$$

$$\mathcal{C}(A) \perp \mathcal{N}(A). \tag{3.45}$$

3.8 Summary

1. The principal components of a vector random variable are an orthonormal basis for the column space of data matrix constructed from samples of the random variable.

2. The SVD provides a method for finding the principal components for data sets consisting of a few observations of many variables, i.e. for cases where $p \ll k$. This is the situation that often occurs in eigenimaging where there are a few images, each of very high resolution.

3. The SVD allows one to find the characteristic roots λ_i of a data set from the singular values s_i. Again, this is useful when $p \ll k$.

4. The number of principal components is the rank of both A and AA^T. Similarly the number of significant principal components is the effective rank ϵ of both A and AA^T.

Chapter IV

A prototype system for classification

This chapter describes a prototype system for using eigenimage processing to classify frontal chest radiographs. The goal of this work was to develop code that could handle the large images used in radiology, and to provide data for analysing how eigenimaging works with respect to chest radiographs. This chapter focuses on implementation with only a minor discussion of the results. The following three chapters analyse and discuss the major aspects of the method, namely image registration, processing of the training set, and processing of test images.

4.1 Aim

The aim of the prototype system is to take a test set containing both normal and abnormal radiographs and classify each test image as either normal or abnormal. The "reference for normal" is a training set of normal radiographs.

4.2 Method overview

The system is based on the face recognition system introduced in chapter 1, and described by Turk and Pentland in their original papers [5,6]. In their system two sets of images are obtained. The first is a training set of "known" face images. The second is a set of test images. Some of these test images are "known" faces while others are "unknown". The known test images are new photos of people already within the training set and the unknown test images are photos of people who are not within the training set at all.

The steps involved in applying this method to chest radiographs were introduced in chapter 1, but they are reviewed again here.

Image registration: Images are spatially aligned and centred with respect to

intensity.

Processing of the training set: PCA is applied to the training set to calculate the eigenimages.

Analysis of test images: Each test image is described as a combination of the weighted eigenimages with one weighting factor per eigenimage. The weighting factors are then used to classify each test image as either normal or abnormal.

For the prototype system the training set consisted of 77 normal frontal radiographs while the test set consisted of 15 normal radiographs and 15 images of pneumonia. Images were all obtained in the manner described in section 1.6. The experiments performed in this and subsequent chapters were all written in MATLAB, a numerical computing environment and programming language [80].

4.3   Image registration

Before eigenimage processing can be applied it is necessary to pre-process all images in both the training and test sets. Pre-processing involves three basic steps: spatial registration, intensity centring, and forming the image into a column vector. Spatial registration and intensity centring are shown in figure 4.1.

As stated in section 1.5.3, image registration is the process of aligning normal structures in the set of images so that variation between images is mostly the result of pathology rather than differences in radiographic technique and normal anatomic variation. For the prototype system registration was performed using a manual affine transformation with both lateral costophrenic angles and both lung apices as landmarks. This technique is similar to the most commonly chosen method for eigenface systems, where a manual affine transform is performed using the mouth and eyes as control points [46]. In chapter 5 the method used here is explained in more detail, and then compared with other registration techniques.

The registration steps for the prototype system are summarised as follows:

1. Four landmark points are manually identified on each image by the author, a qualified radiologist. The points identified comprised both costophrenic angles and both lung apices.

54

| Original | Registration | Centring |
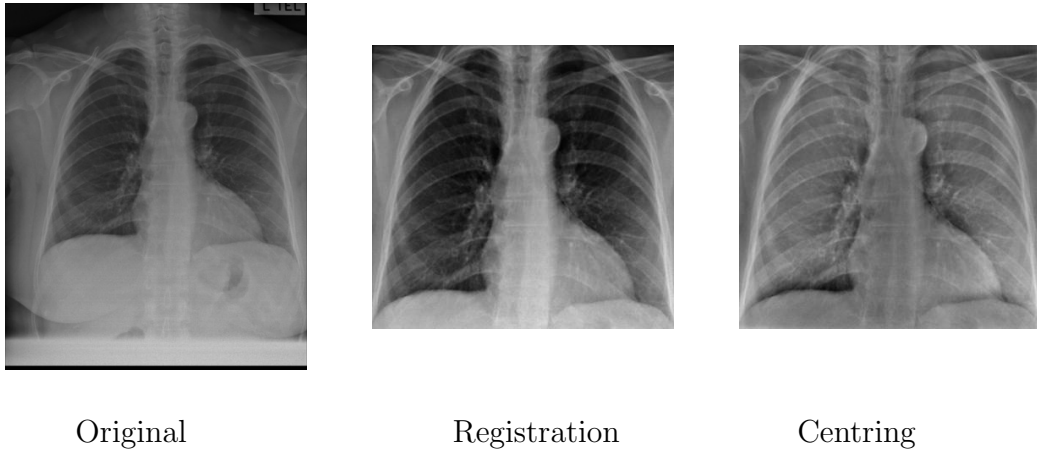
Figure 4.1: **Pre-processing of images:** The image on the left is the original radiograph. The middle image is the result following windowing, affine registration, and cropping. The right image is after intensity centring.

2. The original image is multiplied by a windowing function to reduce the effects of artefacts caused by subsequent cropping. The edges of this windowing function are the identified landmarks, as they lie at the extremes of the chest cavity.

3. An affine transformation is performed on the windowed image so that control points align with those of the average from the untransformed training set. Affine registration allows for the spatial transforms of translation, rotation, scaling, and shearing. The three control points chosen are derived from the landmarks: both costophrenic angles and a point midway between the lung apices.

4. The aligned images are cropped to $m \times n = 1784 \times 1680$ pixels with respect to average control points. This excludes most of the extra-thoracic information while ensuring that the chest cavity is fully represented.

Following registration the images undergo intensity centring. That is, the empirical mean ("average image") from the training set is calculated and then subtracted from all images in both the training set and test set. This step is

referred to as intensity centring as after the subtraction the mean intensity of the training set images is zero. This satisfies the requirement that the data is centred, as explained in section 2.2.3, thus simplifying the PCA process.

After intensity centring all images from both the training set and test set are formed row-wise into vectors, $\mathbf{a}_i$, with each pixel being one element of the vector. Each vector is thus $k \times 1$, where $k = mn \approx 3 \times 10^6$. While this step alters the spatial relationship between the pixels, the order of the elements is arbitrary and unimportant in the succeeding steps, providing it is the same for each image.

For the training set, the image vectors are formed into a data matrix A such that each column of A is one image vector, $\mathbf{a}_i$. Thus A is $k \approx 3 \times 10^6$ rows (pixels) by $p = 77$ columns (images). A is stored on disk as 77 separate files, with each file containing one column of A at 16 bits per pixel.

4.4   Processing the training set

Following pre-processing of the radiographs, PCA is applied to find the principal components of the training set, commonly referred to as eigenimage processing. As presented in section 2.3 the principal components are the eigenvectors of the co-variance matrix $\Sigma = \frac{1}{p-1} A A^T$.

Unfortunately the co-variance matrix Σ is extremely large, with approximately $k^2 = 10^{13}$ elements, and it is too large to be stored on most computers. At 16 bits per pixel it would comprise approximately 50 petabytes of data! However, as presented in section 3.6, it is possible to calculate the 77 principal components and characteristic roots of Σ by studying the SVD of A. This is

$$A = U S V^T. \tag{4.1}$$

The columns of V are the eigenvectors of $A^T A$. This is a square matrix of $p = 77$ rows and columns. Thus, it can easily be stored in a computer's memory. However, the matrix A is around 1 gigabyte of data at 16 bits per pixel. It is therefore necessary to find $C = A^T A$ while keeping A on disk. This is performed doing the calculation one row at a time, reading the rows into memory only as required and keeping a running total of C. Sample code for this is given in figure 4.2. By examining the code it can be seen that the

```
C = zeros( p , p ) ;
for i = 1:k
  a = read_one_row_of(A) ;
  C = C + a' * a ;
end
```

Figure 4.2: **Sample Matlab code for finding $A^T A$:** The *for loop* works by iterating over all k rows of A. It reads the p elements of one row of A by accessing the p files on disk. Then performing the multiplication one row at a time, while keeping a running total, stored as C.

number of multiplications and additions increases linearly with increasing k and quadratically with increasing p.

This process initially took around 15-30 minutes on the computer hardware available near the beginning of this research project. Using hardware available in the Electrical and Computer Engineering Department in 2006, it takes around 3-5 minutes. Further discussion on the speed of computation is presented in section 6.6.

Having found $A^T A$ it is now possible to find both S and V. The diagonal elements s_i of S are the square roots of the non-zero eigenvalues of $A^T A$, while the columns of V are the associated eigenvectors of $A^T A$. To find these eigenvalues and eigenvectors MATLAB's built in function `eig()` was used. This can find the eigenvectors and eigenvalues of a 77^2 matrix in significantly less than a second. The method used by MATLAB for real symmetric matrices is the "DSYEV" routine from the public domain software library "LAPACK" [84].

Having found S and V, it is possible to find the first r columns of U which correspond to non-zero singular values. This was given in equation 3.40, and repeated here for convenience.

$$\mathbf{u}_i \;=\; A\mathbf{v}_i s_i^{-1} \ \text{ for } i = 1 \ldots r \tag{4.2}$$

Alternatively using the economy sized SVD

$$U_r = A\,V_r\,S_r^{-1}. \tag{4.3}$$

Thus, the principal components $\mathbf{u}_i$ are found without ever calculating the complete co-variance matrix Σ.

For speed, the MATLAB code developed for this thesis uses the matrix form given in equation 4.3. The code performs the multiplication in a similar manner to finding $A^T A$. That is, each row of A is read in individually, multiplied by $V_r S_r^{-1}$, then U_r is written to disk one row at a time. Like A, U_r is stored on disk as r separate files each containing one column.

As noted previously, the eigenvectors $\mathbf{u}_i$ are all within the same vector space as the original registered radiographs, therefore it is possible to view them as images. Figure 4.3 shows the first two eigenchests and 4.4 the last two eigenchests respectively. Low order eigenchests represent frequent patterns of variance within the training set, whereas high order eigenchests tend to look more noisy as they represent variance which is less common within the training set.

1st eigenchest $\qquad\qquad\qquad\qquad$ 2nd eigenchest
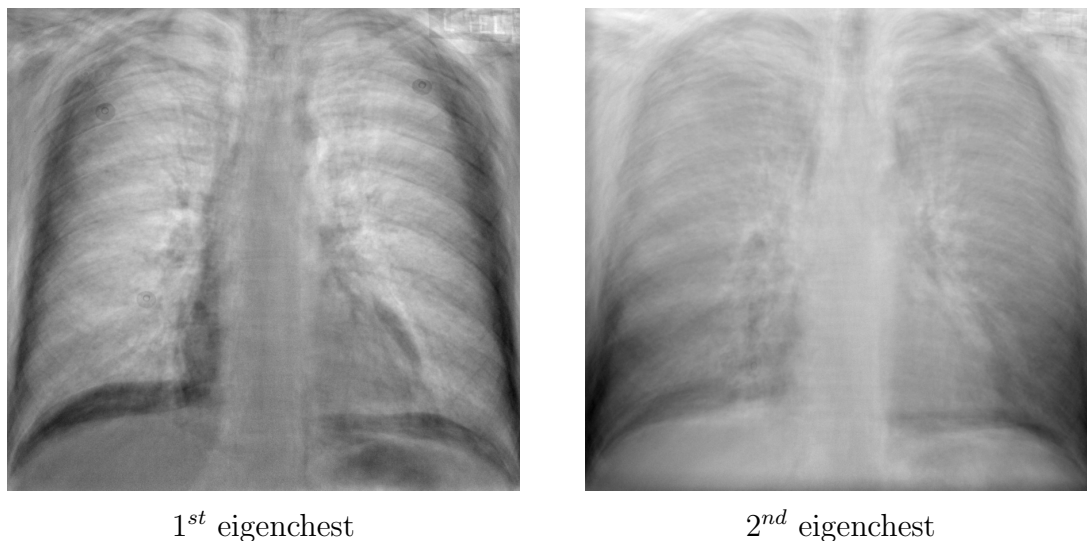
Figure 4.3: **Low order eigenchests:** The left image is the first (lowest order) eigenchest. The right is the second eigenchest. These image represent the variance patterns that are most common within the training set.

75th eigenchest 76th eigenchest
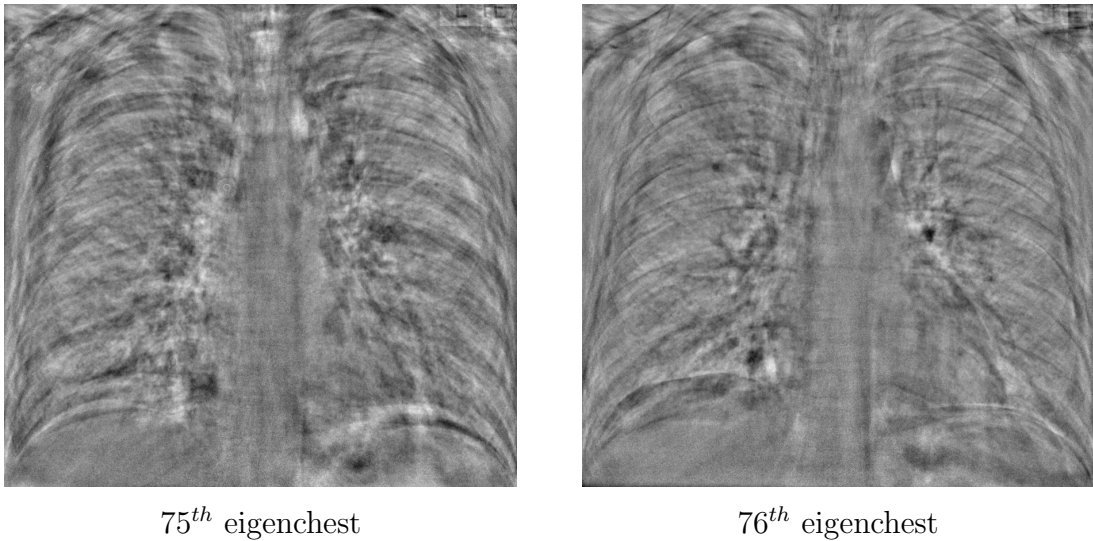
Figure 4.4: **High order eigenchests:** The left is the 75th eigenchest and right is the 76th eigenchest. As expected these is the noisiest image as it represent co-variance patterns that are less common in the training set. It is worth noting that while these image appear dramatic in high contrast, their actual contribution to any test image is small due to their small characteristic roots.

4.4.1 Power spectrum

Once S has been found it is possible to study the power spectrum of the sample co-variance matrix Σ. First the singular values s_i need to be transformed into the characteristic roots of Σ. The relationship is explained in section 3.6.1 and was found to be

$$\lambda_i = \frac{s_i^2}{p-1}. \tag{4.4}$$

Once the characteristic roots λ_i are found they can be displayed as a scree plot, shown in figure 4.5. As expected the low order characteristic roots are much larger than the high order roots. The plot is helpful is showing how much of the total variance is contained within the low order eigenimages.

The significance of the characteristic roots is discussed in detail in chapter 6 which focuses on processing of the training set. That chapter also experiments with varying the number of images within the training set and altering the com-
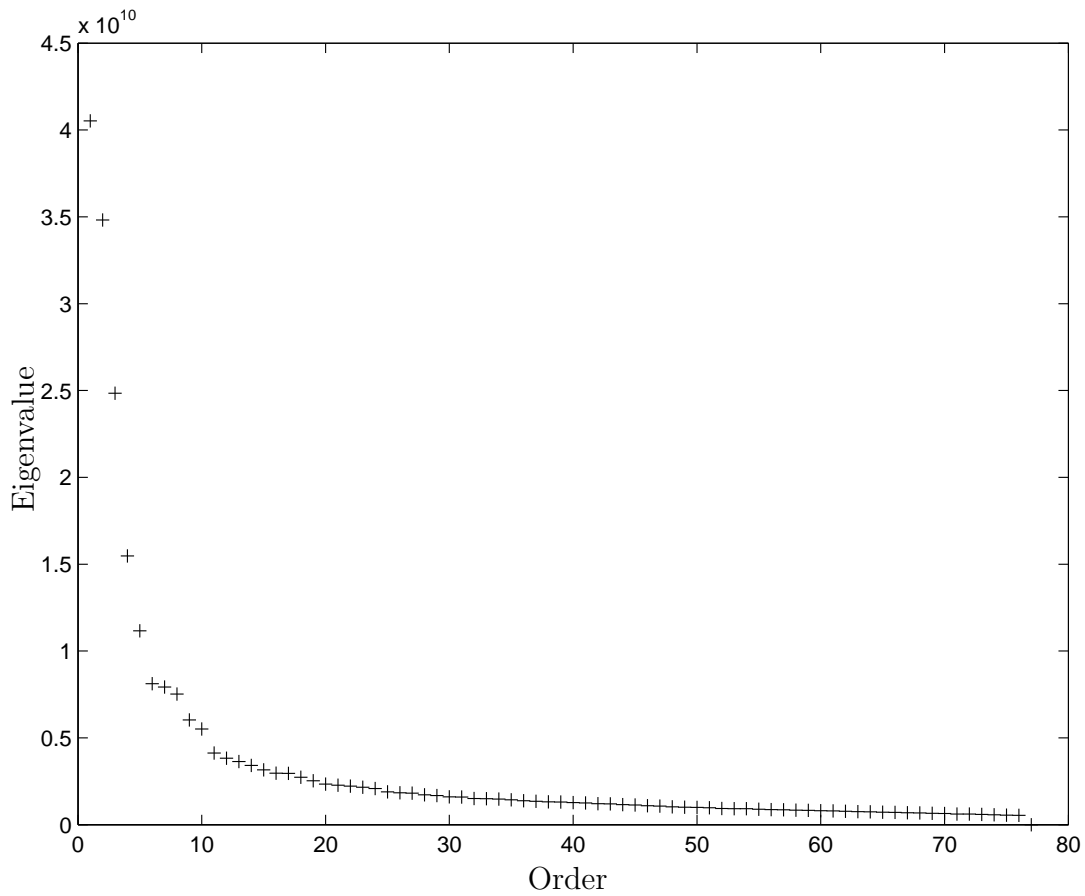
Figure 4.5: **Scree Plot:** The characteristic roots of the training set's co-variance matrix have been plotted. It is called a scree plot because it resembles the accumulation of scree at the base of a steep mountain slope [67].

position of the training set.

4.5 Analysis of test images

Once all the principal components have been found it is possible to approximate each test image using a linear combination of the eigenchests. Parameters derived from this linear combination are used for classification of images into normal or abnormal. Chapter 7 discusses the processing of test images in significantly more detail.

Using Turk and Pentland's method each test image $\mathbf{t}$ is written as a linear combination of eigenchests plus a null space image, thus

$$\mathbf{t} = \sum_{i=1}^{r} w_i \mathbf{u}_i + \mathbf{n}, \tag{4.5}$$

where w_i is the weighting factor associated with the ith eigenchest and $\mathbf{n}$ is the null space image. The null space image is the portion of the test image not represented by the basis of eigenchests (i.e. column space of A) and therefore this portion of the image lies within the left null space of A.

For each eigenimage there is an associated weighting factor, w_i. The weighting factor is found by taking the inner product of the test image and the eigenimage,

$$w_i = \mathbf{u}_i^T \mathbf{t} = <\mathbf{u}_i, \mathbf{t}>. \tag{4.6}$$

Once all the weighting factors $w_1 \ldots w_r$ have been found, it is possible to find the null space image by substitution into equation 4.5. While the original papers of Turk and Pentland acknowledge that the representation of a test image by the eigenimages is only an approximation, they do not find or use the null space image.

4.5.1 Classification of test images

Following calculation of the weighting factors it is usual to use these to classify the test images. For the original face recognition system test images were classified as known or unknown, and for the prototype chest classification system test images are classified into normal or abnormal. The method described by Turk and Pentland used Euclidean length, and therefore this is the method chosen for the prototype system [5, 6]. For each test image $\mathbf{t}$ the weighting factors are calculated. Then the Euclidean length D_E of the weighting factors is found by

$$D_E(\mathbf{t}) = \sqrt{\sum_{j=1}^{p} w_{i,j}^2} \tag{4.7}$$

The Euclidean lengths were calculated for the 15 normal test images and 15 abnormal test images, and are plotted as a cumulative probability function in
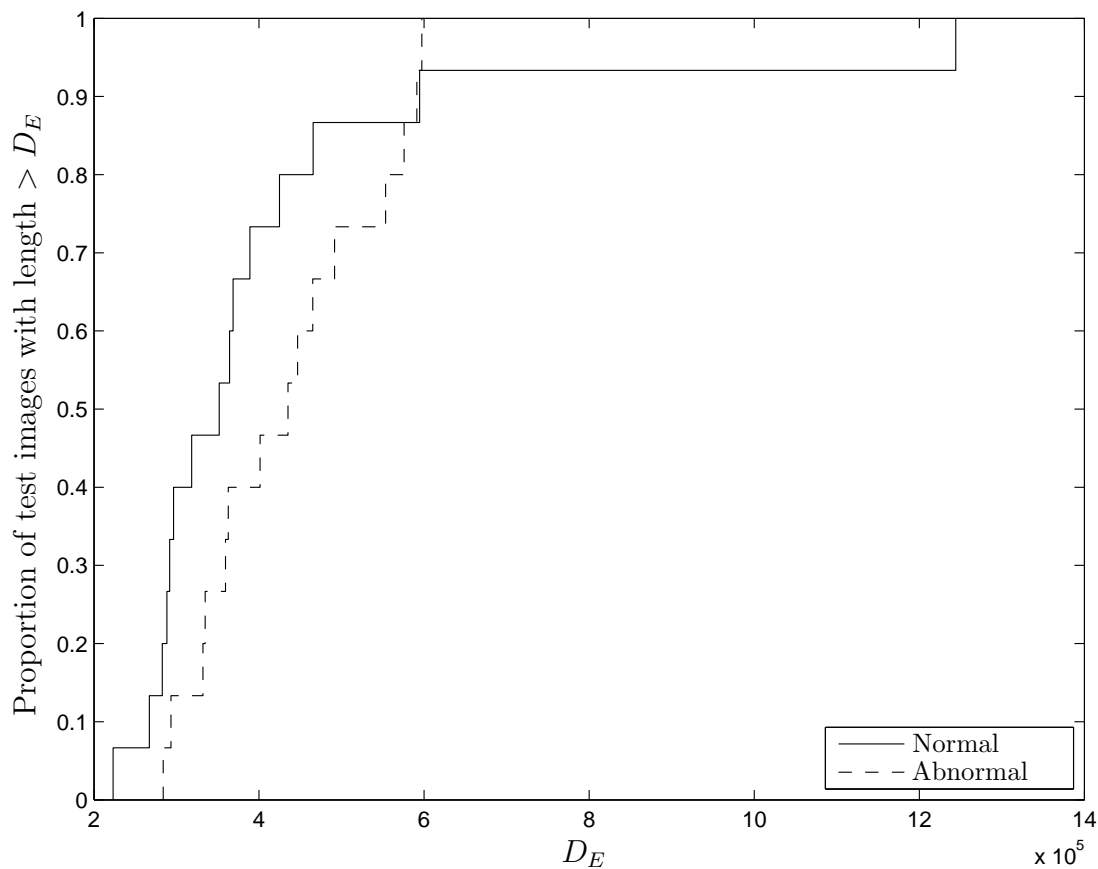
Figure 4.6: **Euclidean length of test image weighting factors:** Cumulative probability function of the D_E for both the normal and abnormal images. As hoped, the lengths for abnormal images tend to be larger than for normal images.

figure 4.6. As expected D_E for the abnormal images tends to be larger than D_E for normal images. In image space, an abnormal test image should be a point away from the cluster of normal training images, and therefore the weighting factors for this test image should be larger. This is similar to face recognition systems where weighting factor vectors for known faces tend to have smaller lengths than unknown faces. A more detailed discussion on weighting factor size is in section 7.1.

Once D_E has been found for each test image it can be used in a simple classification scheme. The method is to classify all images with a Euclidean

length less than a threshold value are classified as normal, while images with a length greater than the threshold value as abnormal.

Using conventional terminology, images that are correctly classified are referred to as either true positives (truly abnormal) or true negatives (truly normal). Similarly, images that are incorrectly classified are referred to as false positives or false negatives. Changing the cut-off at which an image is classified as normal or abnormal changes the number of true positives versus false negatives, and the number of true negatives versus false positives. Sensitivity is the number of true positives, divided by the total number of (true or false) positives. Similarly the specificity is the number of true negatives, divided by the total number of (true or false) negatives.

A receiver operating characteristic (ROC) curve is a common method for displaying the efficacy of a particular test [85, 86]. It plots sensitivity versus specificity for different threshold values between normal and abnormal. For the prototype system the ROC curve is shown in figure 4.7. By studying the curve it can be seen that the Euclidean length classification scheme has predictive value. That is, the test's ROC curve lies above or to the left of the diagonal line. However, all images chosen had been previously reported as normal or abnormal by a fully qualified radiologist, then checked by the author to confirm the diagnosis. It can be assumed that for the images in the test set a radiologist would have a near ideal ROC curve. That is the sensitivity and specificity of the prototype system is significantly worse than a qualified radiologist which suggests that the current system is of limited clinical use without significant modification.

4.6 Checking the experimental code

When developing code for mathematical application it is important to ensure the program is producing correct results. While this is part of the normal debugging process, the methods used to check the system are listed here to aid implementation by others.

After the code runs without errors the following four properties can be checked to ensure that the system produces consistent results.

1. Check that all output matrices are the correct size.
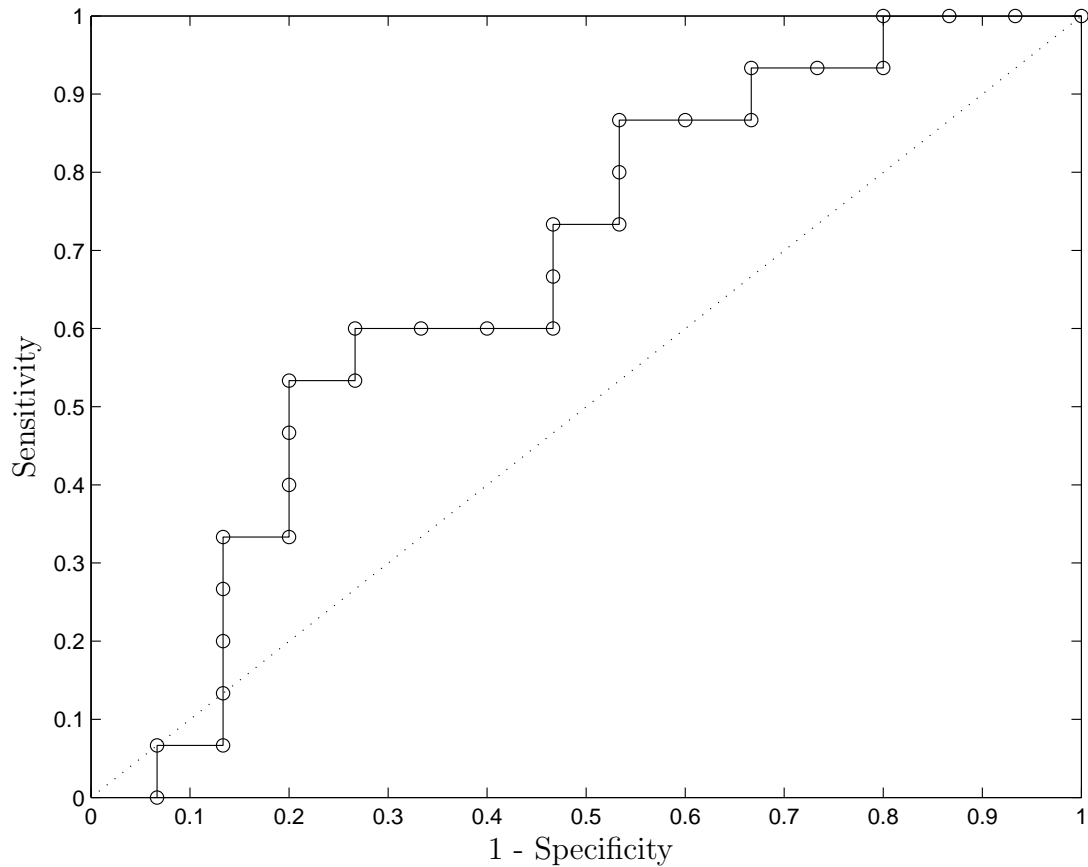
Figure 4.7: **Receiver operating characteristic curve:** This graph plots specificity versus sensitivity. This measures the efficacy of a diagnostic test. A test with no diagnostic value lies along the diagonal, whereas an ideal test follows the upper left corner of the axis. The Euclidean length classification system falls between the two extremes suggesting that it is of diagnostic value.

2. Check that the eigenimages have correct properties.

3. Check that the weighting factors for known test images are those expected.

4. Check that for known test images the variance associated with each principal component is that expected.

Each of these are briefly discussed in the following subsections. These properties also apply to processing performed in subsequent chapters and hence were used for testing that code from those chapters as well.

4.6.1 Size of output matrices

Following the calculations, the system should have produced r eigenimages each of length k. Similarly each weighting factor vector should contain r elements. Intermediate matrices such as S and V should be $p \times p$, with S being diagonal.

For real world data that contains noise it is almost certain that $r = p$. That is, the rank of A should be p. This is because the rank of A is defined as the number of independent columns in A. For such data it is extremely unlikely that any column of A is an exact combination of the other columns. Thus, all columns of A are almost certainly independent. However, the effective rank ϵ may be less than p (see section 3.5).

4.6.2 Properties of eigenimages

Each eigenimage is calculated indirectly from A, via V and S, as explained in section 4.4. According to the SVD, U should be an orthogonal matrix, implying that its columns are orthonormal. Thus

$$\mathbf{u}_i^T \mathbf{u}_j = \begin{cases} 0 & \text{for } i \neq j \\ 1 & \text{for } i = j \end{cases}. \tag{4.8}$$

For the prototype system this was indeed found to be the case within the precision of the numerical computations.

4.6.3   Weighting factors of known images

One special case is solving equation 4.5 for a test image which is part of the training set. Specifically this is when $\mathbf{t} = \mathbf{a}_i$. The weighting factors for the ith training image are

$$
\begin{bmatrix} w_1 \\ \vdots \\ w_r \end{bmatrix} = U_r^T \mathbf{a}_i \tag{4.9}
$$

Using the economy sized SVD in equation 3.23 it is known

$$
\begin{aligned}
U_r^T A &= S_r V_r \tag{4.10} \\
\Rightarrow U_r^T \mathbf{a}_i &= S_r \mathbf{v}_i \tag{4.11}
\end{aligned}
$$

Thus if the weighting factors of the ith training image are found then they should be exactly $S_r \mathbf{v}_i$. Since S_r and V_r are already known this provides a convenient method to check that the code for calculating the weighting factors provides the expected results.

4.6.4   Variance of known images

It is known that the variance associated with an eigenimage is given by the characteristic root. It is also known that using the training set images as test images, as in the previous section, there are $p = 77$ measurements of the weighting factor associated with an eigenimage. Thus there are two ways to calculated the variance associated with a particular eigenimage.

1. From the eigenvalue of the sample co-variance matrix, $\lambda_i = \frac{s_i^2}{p-1}$.

2. From weighting factors found using the training images as test images, $\mathrm{var}[w_{i,1} \ldots w_{i,p}]$.

Since the data used in both methods is identical, the results should be identical if there is no error in the MATLAB code. For code developed for the prototype system, both methods produced identical results (within numerical precision).

4.7 Conclusion

The prototype system demonstrates that technique used by Turk and Pentland in their face classification system can be successfully applied to radiographs. This work was presented at two New Zealand radiological meetings, one international radiological meeting, and one international computer engineering meeting [87–90].

The first conclusion from this work is that the method can be applied to the large images found within radiology departments. The application of eigenimage processing to radiological images is novel, and therefore a successful implementation is significant.

The second conclusion is that the classification system has predictive value, although the predictive value of the system is less than that of a skilled radiologist. For other image types people have used eigenimage processing to find the weighting factors, but then used different classification schemes to achieve better results [55, 57, 58, 91]. However, different classification schemes have different advantages and disadvantages depending on the application.

For the prototype system some arbitrary decisions were made. Firstly, the manual affine transformation was used as a convenient and fast registration system. There are many other manual and automatic registration systems available and it is necessary to evaluate the efficacy of these.

Secondly the training set included only 77 normal radiographs. The number of images in the training set is likely to affect the usefulness of the system and this needs to be tested. There is also the possibility of including abnormal images within the training set as this will allow the eigenimages to better represent pathology. This possibility is discussed further in sections 6.3.4 and 6.7.

For evaluating the prototype system there was a limited number of test images, with only one form of pathology (pneumonia). It will be important to evaluate the system with a larger test set which includes a wider range of pathology. This will provide insight on which pathologies are best detected with this technique.

Now that it has been established that eigenimaging can be applied to radiographs it is possible to consider developing a range of clinically useful applications. As well as CAD, other applications of eigenimage processing have included

image compression and image enhancement, as discussed in section 1.5.

4.8 Summary

1. It is possible to apply eigenimage processing to frontal chest radiographs.

2. Classification is achievable, even if clinical efficacy is currently limited.

3. It is necessary to study the effects of using different registration systems.

4. It is necessary to study the effect of the size and composition of the training set.

5. It would be valuable to study the processing of test images. In particular, it would be helpful to attempt classification using more sophisticated metrics than Euclidean length. Also, it is important to consider the information not represented by the eigenimages (i.e. the null space image).

6. The test set needs to include more images with a wider range of pathologies.

Chapter V

Registration

The aim of the experiments described in this chapter is to assess the effect of different registration systems on eigenimage processing. As stated earlier, image registration is the process of aligning two images with respect to normal structures. In eigenimage processing, registration is applied to all images in both the training and test sets.

The goal of registration is to remove variance that arises from patient positioning and normal anatomical variation, while preserving variation that aids in diagnosis. That is, registration attempts to remove factors such as image rotation, exposure, magnification, patient position, and patient size.

Image registration is also important elsewhere in medical image processing, for example in digital subtraction angiography, temporal subtraction, image fusion (e.g. PET-CT), and image guided surgery. Hundreds of articles, books, and theses have been published on the subject with Pluim and Fitzpatrick finding over 200 articles in 2003 alone [92]. Good review articles can be found in both the engineering and medical literature [92–98]. Unfortunately most of the research into medical image registration has been in the field of 3D neuroimaging, with significantly less work published on chest radiograph registration.

This chapter introduces commonly accepted terms used to describe a registration system. Following this, there is a short review of several common methods used in medical image processing, where it is found that all automatic methods for registering chest radiographs have significant flaws. Thus, four semi-automatic methods were selected for evaluation in conjunction with eigenimage processing. Experiments performed in this chapter examine the effect each registration routine has on the average image, training set variance, and accuracy of classification.

For all other chapters of this thesis, registration was performed using a rigid

global affine transformation based on manually selected control points. As noted in the introduction, section 1.6, all images have their intensity manually scaled by the radiographer. If this were not the case the registration system would also need to remove variations in image exposure.

The work contained in this chapter was presented at an international optical engineering conference [99].

5.1 Taxonomy of medical image registration

When discussing medical image registration it is important to accurately describe both the alignment problem and the registration method used to solve it. An article by van der Elsen presents a classification scheme for describing registration techniques [100]. This classification scheme is widely used and provides a simple method to describe medical registration systems. An overview of the scheme is given below.

5.1.1 Source and target images

When describing alignment problems it is necessary to describe what is to be aligned. By convention, a source image is aligned to a target by altering the source image. Often both the source and target are images, although in some situations the target may be a set of co-ordinates within the image's domain. In this chapter the target for automatic routines is a selected normal image. For all other chapters, and the semi-automatic routines in this chapter, the target is a set of co-ordinates for the average control points. MATLAB's registration routines refer to the source and target as "input" and "base" respectively.

5.1.2 Dimensionality

Problems are described in term of the dimensionality of both the source and target images. For example, the problem could involve registering a 2D frontal chest image, onto a 3D chest CT. Registration systems exist for various combinations of the source and target being 2D, 3D, or time varying data. For this thesis, registration is always of a 2D source image onto a 2D target.

5.1.3 Origin of image properties

Registration may be based on data intrinsic to the image or data extrinsic to the image. Intrinsic information is data within the image, such as bony landmarks. Extrinsic information is data not found within the image, such as table position for CT images. This thesis uses intrinsic information, as only minimal extrinsic information regarding the subject's position is available.

5.1.4 Domain of transformation

This describes how the registration technique is applied. The term "local" is used to describe how a small portion of the image is transformed, while "global" is used to describe how the entire image is transformed. In a registration system the local transformation can differ from the global transformation. Periaswamy's method, section 5.2.4, is an example in which the local transform is different from the global transform.

5.1.5 Nature of the transformation

This describes how the data is transformed. Common terms include rigid, linear conformal, affine, projective, curved, deformable, and elastic.

Rigid: These transformations only allow for translation, rotation, and reflection. The underlying image is otherwise unaltered. Specifically, the distance between any two points is unaltered in the source and registered image. Under this transformation a square will remain a square of the same size.

Linear conformal: This is similar to rigid registration except images can be scaled. Angles remain the same. Thus a square remains a square, but possibly of a different size.

Affine: This allows for translation, rotation, scaling, and shearing. In this technique parallel lines remain parallel. A square may be transformed into a parallelogram.

Projective: This is defined as a transformation in which straight lines remain straight, but parallel lines may become divergent. This provides for a wide

range of transformations with a square becoming any convex quadrilateral.

Curved, deformable, or elastic: These techniques allow a straight line to be bent during the registration process. Synonyms include deformable registration and image warping. Registration models must define how much deformation is allowed. Models which are globally elastic may be locally inelastic. Under these transformations a square may become any shape.

Many registration systems use a combination of techniques. For example, the method described by Periaswamy is "globally elastic and local affine" [96]. This means that over short distances within an image the transformation is affine but different parameters for the affine transformation are used across the whole image, leading to a globally elastic transformation.

5.1.6 User interaction

Systems require varying levels of user input. Systems are often classified as "interactive", "semi-automatic", and "automatic". While it is difficult to objectively quantify the level of user interaction, some systems require more than others. An ideal system would be fully automatic and require no user interaction. However such systems are very difficult to design and it is common to require some user input.

5.1.7 Parameter determination and error metric

When registering two images the parameters for the transformation need to be determined. Sometimes this is done explicitly from control points, or alternatively it can be done using a metric quantifying "how unregistered two images are".

The affine registration used in the prototype system, chapter 4, was performed using manually selected control points to determine the parameters of the transformation. Other methods may involve automatically finding control points to determine the transformation parameters. Lehmann describes a system where landmarks are automatically found within medical images [101]. Similarly, Zhao describes a system for chest radiographs where edges are detected and aligned [102].

For many automatic systems the parameters of the transform are based on an error metric relating to how well registered any two images are. An error metric allows for search systems to find the lowest error (i.e. best registration). There are many error metrics used, but probably the most common is the mean squared error. This is the mean of the squares of point wise difference in intensities between image pairs. This metric is simple and fast to compute.

Two other common error metrics are the ratio of image uniformity and mutual information. The ratio of image uniformity is the normalised standard deviation of the ratio of image intensities. This method has the advantage of not requiring the overall image intensity to be the same, but tends to lead to more difficult optimisation problems. The mutual information metric is useful as it implicitly handles variations in contrast and brightness although it is a very computationally expensive method.

5.1.8   Multiple resolution approaches

In automatic registration systems it is often useful to perform the registration at different scales. That is, images are aligned at a large scale (coarse alignment) and then progressively aligned at smaller scale (fine alignment). Several methods can be used for multiresolution systems including Gaussian pyramids and wavelet methods [4, 96, 103–106]. Multiresolution methods have several advantages including:

- They tend to be faster because coarse alignment can have a larger step size.

- They tend to avoid local minima.

- Registration parameters can be different at different scales. For example, at a coarse scale the registration can be more elastic than at a fine scale.

5.2   Review of common methods and existing software

The aim of this chapter is to assess the effect of different registration systems on eigenimage processing. As such, it was decided to take well understood methods and apply them, rather than to develop and implement new methods. Un-

fortunately there are no widely accepted routines developed for automatically registering chest radiographs.

The lack of any commonly accepted routines most likely reflects the difficulty in registering chest radiographs. Difficulties include the fact that radiographs are projection images in which overlapping features need to be transformed in different ways. For example in chest radiographs the anterior ribs might need to translate to the left, while the overlapping posterior spine might need to translate to the right. In this conflicting situation many algorithms produce very distorted images. Automatically registering chest radiographs may be problematic because any pathology present within a source image can make registering to a normal target image unreliable. For example a large mass in the left side of the chest could erroneously be registered so that it aligns with the heart in other images.

Since there are no widely accepted routines for registering chest radiographs, several general routines and software packages designed for medical images were tested. Methods reviewed in this chapter include MATLAB's semi-automatic routines, a general purpose routine (Lucas-Kanade), and multipurpose medical packages (AIR5 and Periaswamy's routines). Of the automatic methods (Lucas-Kanade, AIR5, and Periaswamy's method), none worked sufficiently well to make testing in conjunction with eigenimage processing worthwhile. However, the problems encountered illustrate the difficulties in any automatic registration system.

5.2.1 MATLAB's routines

MATLAB's Image Processing Toolbox contains several routines for image registration. They are all semi-automatic systems based on manually selected control points [79, 80, 107]. Routines include linear conformal, affine, and projective. These are simple, fast, and well documented. Their effect on eigenimage processing is investigated in section 5.3.

5.2.2 Lucas-Kanade

This method was first proposed by Lucas and Kanade in 1981, and is a fully automated system that is has found wide use in non-medical image registration and optical flow problems. The method involves finding a difference image be-

tween the source and target. The source image is then warped to minimise this difference image. A steepest descent algorithm is used to iteratively find the best warp.

There are many variations on the method. Baker *et al.* have written two good review articles on gradient descent algorithms and their subsequent articles provide a unifying framework for many of the variations [108–114]. Their articles and five technical reports are all available on the web and include working Matlab source code [115]. However, this publicly available code was investigated but found to be extremely slow and unworkable for even moderate sized radiographs.

5.2.3 AIR

"Automatic Image Registration" tools are a free set of image registration routines for use with medical images. Version 1 of the software was used for registration of PET images [116]. Version 5, the latest version, is available free of charge on the world wide web [117].

The package can be used for:

- Intrasubject, intermodality registration. e.g. PET to MRI

- Intersubject, intramodality registration. e.g. MRI to MRI

- Subject to atlas registration.

The package provides for a variety of linear and non-linear models with multiple error metrics available. Data interpolation can also be performed using a variety of methods. While the software is normally used for 3D data sets, it has also been validated for registration of 2D brain MRI slices. It has not been validated for chest radiographs.

The AIR5 routines are written in "C" and use a file format known as "Analyse-7.5" or "SPM". Exporting images from Matlab into this format can be achieved using `MRI_toolbox-1.5` available as part of the open source "Bioelectromagnetism Matlab Toolbox" [118].

On attempting to use this software for chest radiographs it became clear that the intended use is for brain image registration, where images tend to be CT, MRI, or PET. These classes of images are very different from chest radiographs,

both in terms of both pixel/voxel size and intensity range. More importantly, what is a reasonable deformation to align two brain images is quite different from what is reasonable when aligning two chest radiographs, and it was not possible to get sensible results using the AIR5 routines.

5.2.4   Periaswamy's method

Periaswamy designed a general purpose system for deformable automatic medical image registration [96, 119, 120]. The system is based on a multiscale approach using Gaussian pyramids. Each transform is locally affine, but has a global smoothing constraint. The error metric he employs is the mean squared error.

His Ph.D. thesis clearly describes the system and how to tune it to various image types. The MATLAB source code is available on the web, and the code is well documented and performs well. He extends his method to provide reasonable results with intensity variations across the image as well as missing data in the image [121]. Background intensity gradients are common in MRI images, as signal intensity tends to drop off further away from the receiver coil, but are uncommon in radiographs.

Using the source code provided on the internet it was possible to evaluate the technique. His code has many tunable parameters to suit different registration problems. For example, it was possible to set parameters so that the routine assumes there is no background intensity gradient within the image.

However there were problems with the computation time and the memory requirement of his routines. Within his thesis he primarily focuses on brain MRI images which are much smaller images than high resolution chest radiographs. To overcome these processing limitations, the chest radiographs were downsampled using MATLAB's imresize() function with bilinear interpolation. Images were downsampled to 10% of their original width and height, giving just 1% of the original number of pixels (205×204). Despite this, the routine took approximately fifteen minutes per image.

Figure 5.1 is an example of registering a normal source radiograph onto a normal target image. The areas outside the thoracic cavity required significant warping in order to perform the registration.

An important task for registration in eigenimage processing is to align an
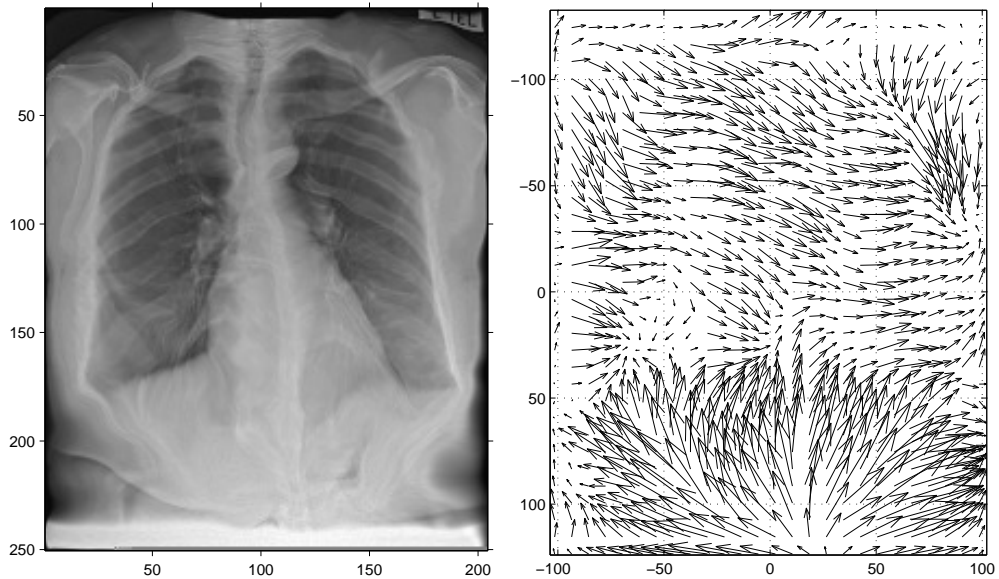
Figure 5.1: **Automatic deformable registration:** Example of a normal image undergoing a deformable registration process (Periaswamy's method). The technique is a mesh of small scale affine transformations, but is globally smooth. The left image is post registration, while the right image is the optical flow used to transform the source image.

abnormal image to a normal target image. This situation arises when the test image to be classified contains pathology. Figure 5.2 displays the result registering an abnormal source image (right lower lobe pneumonia) onto a normal target image. In this example the source image (not shown) had a large area of increased pixel intensity at the right base. In order for the registration process to align the abnormal source image to a normal target image it had to deform the source image very significantly. The extreme deformation is visible in the chest wall.

Based on these initial attempts at using Periaswamy's deformable registration routines several choices were made to improve the quality of registration for chest radiographs.

Pre-processing of Image: It was clear that the deformable registration method led to unexpected results outside the chest wall. For this reason each test image was windowed, translated, and cropped based on the manually se-
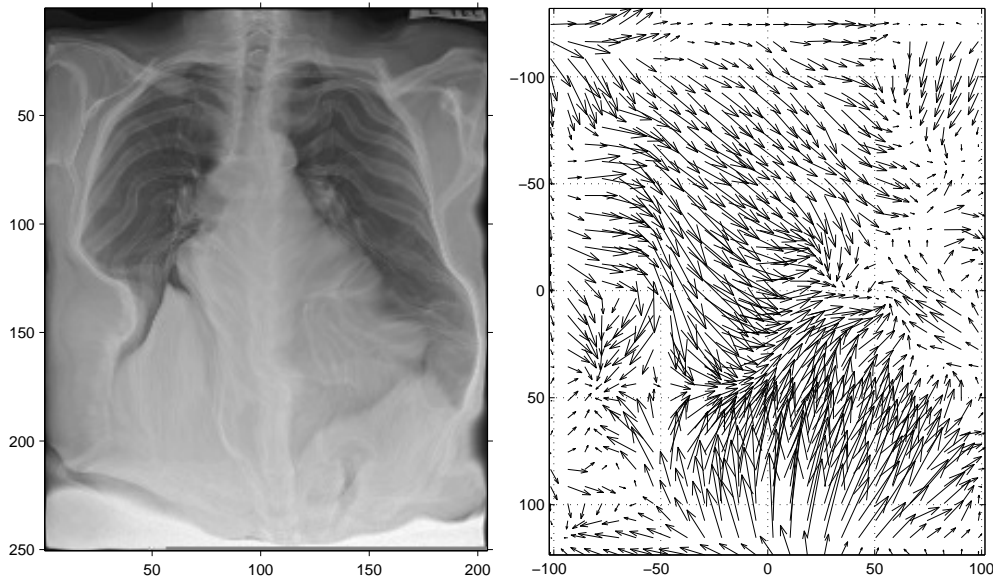
Figure 5.2: **Automatic deformable registration 2:** The left image is a right lower lobe pneumonia following registration (Periaswamy's method) onto a normal target image. The right image is the optical flow used to transform the source image. The presence of pneumonia causes the registration process to extensively deform the chest in order to match the abnormal source image onto a normal target image.

lected landmarks. This made the source images identical to those described in section 5.3.3 and demonstrated in figure 5.4. While this pre-processing no longer makes the system fully automatic it does reduce the effect of extra-thoracic data, which is assumed to be extraneous to the classification problem.

Minimum Pyramid size: The method uses a multi-resolution approach where the image is registered from coarse to fine. The minimum pyramid size is the width of the smallest resolution. For Periaswamy's work on brain registration it was reasonable to allow for very small deformation. Choosing a larger the minimum size means that the registration is only done at larger scales, thus restricting small deformations. This also significantly reduces the computation time.

Width of the box used for the least squares estimate: Increasing this value

alters the concept of "local," allowing so-called local warpings over a wider spatial area.

Following these changes to the parameters it was possible to decrease the processing time to about eight minutes for a very small (219×219 pixel) image. Problems with chest wall distortion were reduced, but persisted as a significant problem.

5.3   Methods assessed in conjunction with eigenimage processing

All of the automatic methods assessed were found to be both unreliable and slow for chest radiographs. In particular the methods tended to fail when pathology was present. Hence, four reliable and fast semi-automatic methods for image registration were chosen to be tested in conjunction with eigenimage processing. The four selected routines have a wide variety of properties and should therefore provide insight into the effect of registration on eigenimage processing.

This section provides details on how each of the four semi-automatic methods were performed. The methods are all based on MATLAB's routines. For all semi-automatic methods the steps involved are:

1. Manually identify landmarks.

2. Window the area outside the chest cavity.

3. Apply one of the four registration routines.

For convenience the methods are labelled: translate, linear conformal, affine, and random. It is presumed that the order of poorest registered to best registered is: random, translate, linear conformal, affine.

5.3.1   Identification of landmarks

For the research presented in this thesis, four landmarks were identified when each chest radiograph was imported into MATLAB: both costophrenic angles and both lung apices. The points were identified by the author, a fully qualified radiologist. An example of the landmarks is given in figure 5.3. These landmarks

are used for both determining the size of the chest cavity and then for subsequent semi-automatic registration routines.

5.3.2 Windowing

Most chest radiographs have a large portion of the image representing structure outside the chest cavity. This information is extraneous to the area of interest on the radiograph. It was therefore desirable to remove this extra-thoracic area. This was done using a technique known as windowing. Areas outside the chest cavity (as determined by the landmarks) have their pixel intensity value set to zero. Artefacts are common when an image undergoes cropping followed by Fourier analysis. It is unknown if similar artefacts occur when an image is cropped and the following processing is PCA. However, to reduce possible artefacts a smooth cosine curve of 75 pixels width was used as a transition. Figure 5.3 has an example of an image pre- and post- windowing.
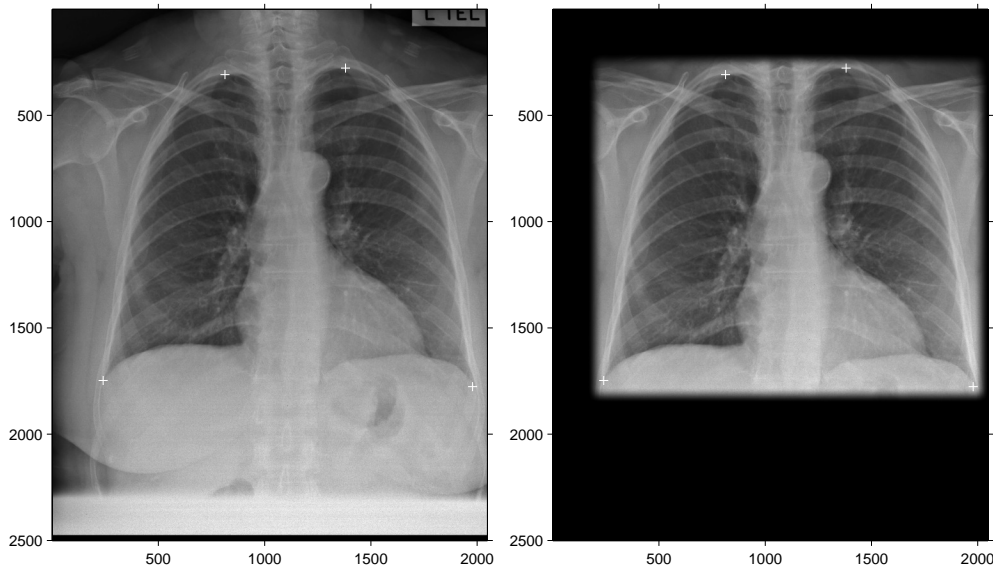
Figure 5.3: **Windowing of an image:** The left image is a source image with the four landmark points identified by white crosses. The right image is the same radiograph with the window routine applied so that extra-thoracic information is minimised.

5.3.3 Translation only registration

The "translate" method of registration allowed only translation of the source image. No rotation or scaling was permitted. The magnitude of translation in both the horizontal and vertical direction was controlled by the manually selected landmarks. The images were then cropped so that the image frame was the smallest bounding box that fitted all the landmarks from the training set.

Specifically the steps involved were:

1. The source image was windowed around the landmarks, as per section 5.3.2.

2. The centre of the landmarks was found. The vertical centre was defined as halfway between the uppermost lung apex and the lowermost costophrenic angle. The horizontal centre was defined as halfway between the right costophrenic angle and the left costophrenic angle.

3. The source image was then translated so that the centre of its landmarks lie on the origin in image space.

4. The image was then cropped to remove extraneous information. The size of the bounding box was the smallest height and width that contained all the landmarks from the training set images. This cropping ensured that all images were of the same size despite the chest filling a different fraction of the image.

The translation was performed using MATLAB's Image Processing Toolbox `imtransform()` function with bilinear interpolation.

Two examples of translated images are given in figure 5.4. In these examples the bounding frame of the image is much larger than the chest cavity of either image. This is because the bounding frame must be large enough to accommodate both the tallest and widest of the chest images.

5.3.4 Linear conformal and affine registration

The "linear conformal" and "affine" methods were the two other registration routines assessed in conjunction with eigenimage processing. Again, both are
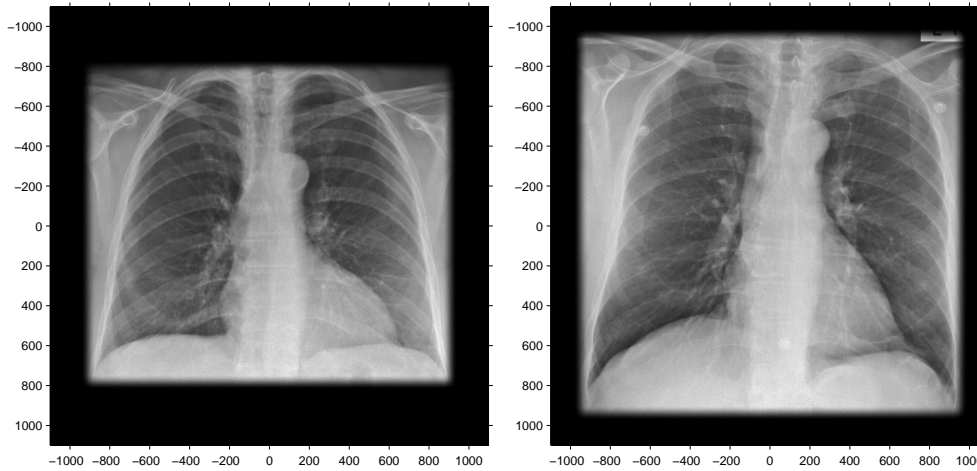
Figure 5.4: **Translation of images:** Example of two images after translation. The centre of all the landmarks is now at the origin in image space (note the image axes). The image's bounding frame is larger than either of the chest cavities shown because some chest cavities are much larger than the two examples shown here.

in MATLAB's Image processing toolbox. These forms of registration map prese-lected control points on the source image to corresponding target control points.

Linear conformal transformations allow for translation, magnification, and rotation of the source image. Affine transformations allow for translation, mag-nification, rotation, and shearing of the image. Again the translation was per-formed using MATLAB's Image Processing Toolbox `imtransform()` function with bilinear interpolation. The steps for both transformations are similar:

1. The image was windowed around the chest cavity, as per section 5.3.2.

2. The source image's three control points were derived from the previously selected landmarks, and comprised both costophrenic angle landmarks and a point midway between the lung apex landmarks.

3. The target image's control points were the geometric centre of each control point from the training set. i.e. the source image's left costophrenic angle's control point was mapped onto the average location of the costophrenic angle.

82
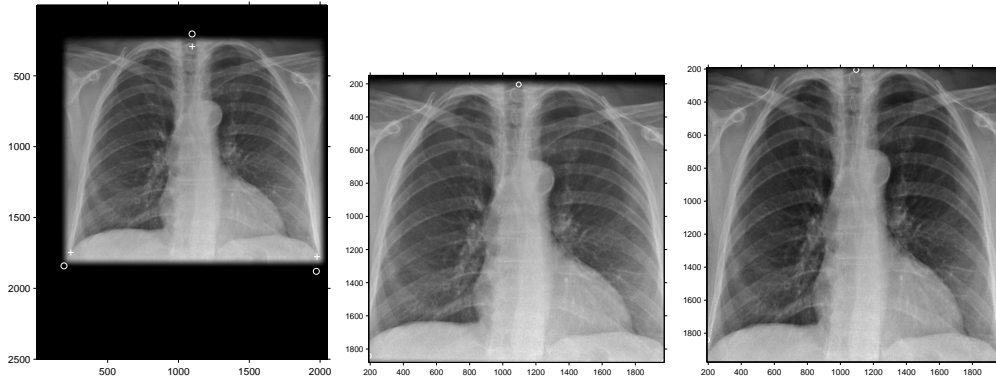
Figure 5.5: **Linear conformal and affine transformations:** The image on the left is post-windowing but pre-registration (i.e. the source image). The source image's control points are marked with a cross, while the target control points are marked with a circle. The centre image is the radiograph after linear conformal registration. The image on the right is after affine registration.

4. Images were transformed using either an affine or linear conformal routine from MATLAB's image processing toolbox.

5. Images were then cropped around the translated landmarks.

Figure 5.5 demonstrates one image before registration and the same image after linear conformal registration and global affine registration.

5.3.5 Random orientation

For experimental purposes a form of registration was developed, labelled "random," where the images where deliberately made less registered than the starting image. The method involved starting with the windowed and translated images from section 5.3.3, then deliberately randomly rotating and translating the image. This random orientation technique provided a simulated example of poorly performed registration. Specifically the steps involved were:

1. The starting images were the post windowing, post translation images from section 5.3.3

2. Each image was randomly rotated. This was done using MATLAB's Image Processing Toolbox's `imrotate()` function using bilinear interpolation. The angle of rotation was randomly chosen from a Gaussian distribution with a mean of 0° and a standard deviation of 20°.

3. Each image was translated. Again, this was done using MATLAB's built-in functions. The x and y translations were calculated independently. Their magnitude was randomly chosen from a Gaussian distribution with a mean of 0 pixels and a standard deviation of 40 pixels.

5.4 Measuring the quality of registration

Having determined four methods of registration to test in conjunction with eigen-image processing it is necessary to have some metric for comparing the quality of registration. The two indirect methods chosen for measuring registration success were: blur of the average image and training set variance. These indirect methods were compared to the one direct method for measuring registration success: accuracy of classification.

5.4.1 Blur of the average image

As discussed in section 4.3, following registration it is necessary to find the empirical mean of the training set, i.e. the "average chest". Figure 5.6 shows the average chest from the training set following the four forms of semi-automatic registration routines.

As expected the average images are very blurry. The conjecture is that a set of well registered images should produce a sharper average image. Conversely a set of poorly registered images should produce an average image with more blur. Therefore, comparing the blur in the average images should give an indication of how well registered the images are. To enable the comparison a metric for this blur is required.

Within an image, blur can come from a number of sources such as the image capture device, movement, and poor focus. Most measurements of blur are performed with the ultimate aim of deblurring an image. Often this is done by estimating the PSF (point spread function) or the MTF (modulation transfer

84

function). This usually involves modelling the image capture device, or having an estimate of a sharp image [122–125].

Figure 5.7 shows the intensity profile for a vertical strip of each of the four average images. To make direct comparison possible, average images from the translate and random routines were cropped to be the same size as the average images from the linear conformal and affine routines. The intensity profiles demonstrate that edges within all four average images are often wide and poorly defined.
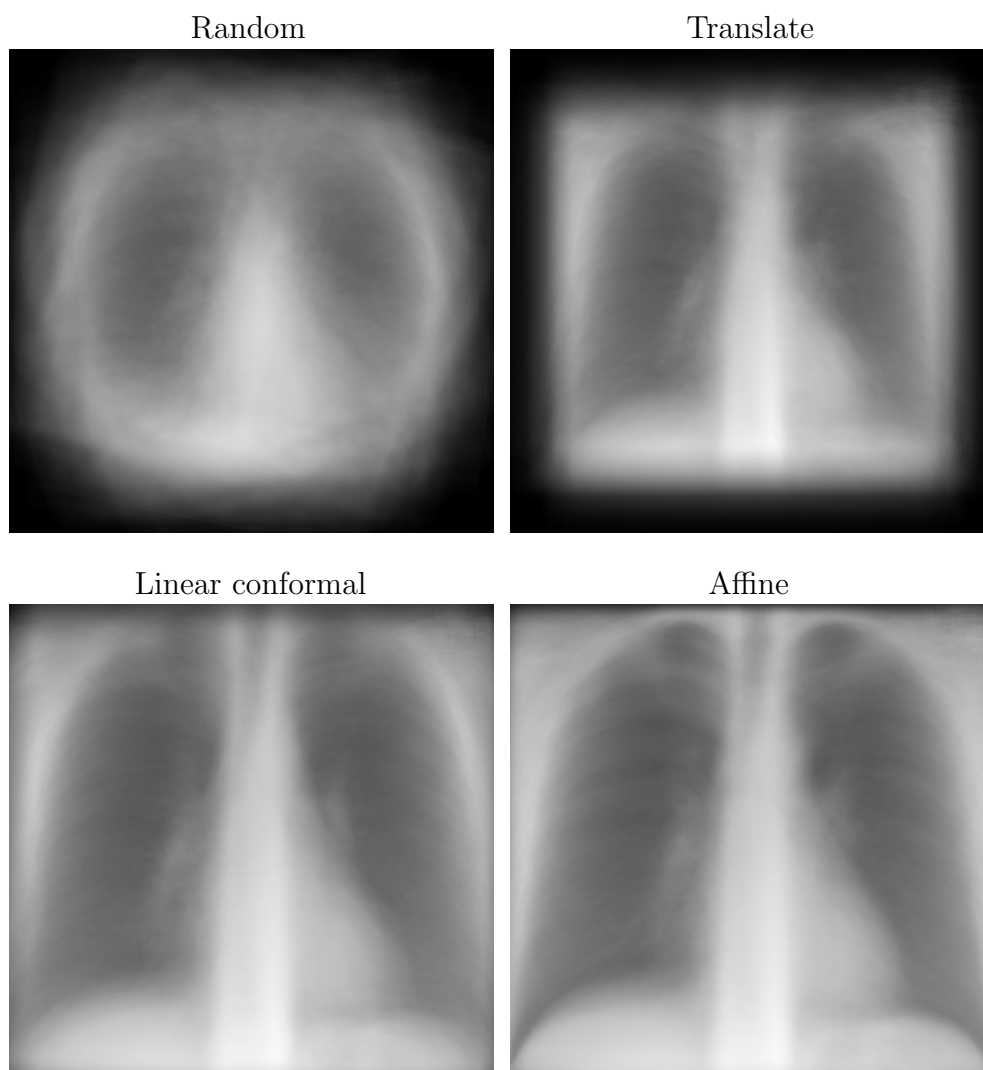
Figure 5.6: **Comparison of average chests:** The empirical mean of the training set following the four different forms of registration.
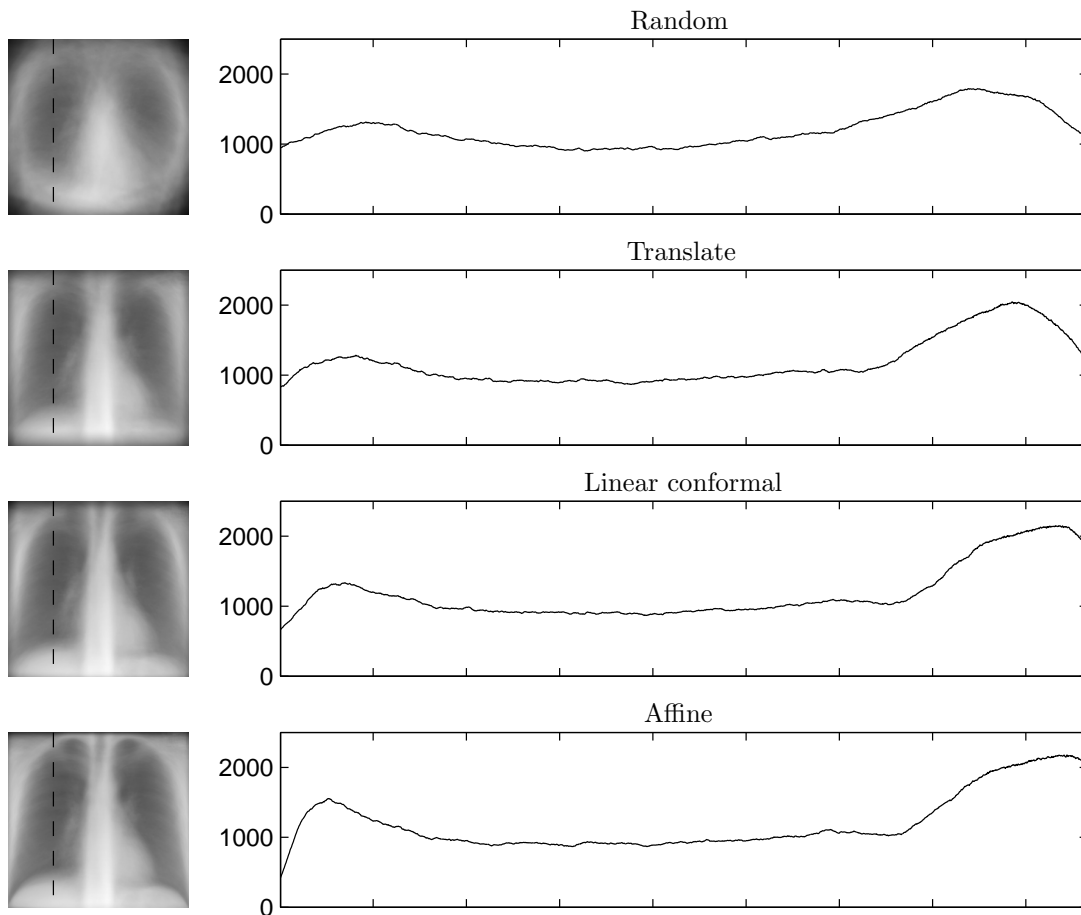
Figure 5.7: **Profile of average chest images:** The four graphs show the intensity profile of a vertical strip from each of the average images. The left end of each profile corresponds to top of the image, while right end of each profile corresponds to the bottom of the image. It can be seen that to measure blur by identifying edges is extremely difficult due to the soft (wide) nature of the edges.

Within average images, blur predominantly arises from registration artefact. There is no camera device to model, nor any sharp reference image to compare to. To compare the blur in the four average images a "no-reference metric" is required, i.e. one that does not need a camera model or a reference image.

One no-reference blur metric is to measure the width of edges within an image [126, 127]. Both Canny and Sobel edge detection filters were used to identify edges. However due to the very smooth nature of the edges neither filter

could reliably detect the edges, making measurement of the width of the edges unreliable. Thus, the metric was not useful.

Figure 5.8 displays the intensity histograms for the central region of each average image. It can be seen that the range of intensities varies between the four forms of registration. This was expected because blurry images often have less contrast than sharp images.

There are many ways to measure the spread of pixel intensities within an image. Three common methods are: image variance, interquartile range (IQR), and image entropy. IQR is the range from the 25 percentile to 75 percentile. Image entropy is measurement of randomness of the intensities and is calculated as

$$\xi = -\sum p \log_2(p) \tag{5.1}$$

where p is the number of pixels at any given intensity [107].

These results for these three metrics are presented in table 5.1, but unfortunately none of the metrics seem useful. While all place the randomly registered image as being the most blurry, they are not consistent with respect to which average image is least blurry. The most useful metric is the average image variance which places all four methods in the expected order.

Registration Method	Variance	Entropy (ξ)	IQR
Random	1.19×10^5	6.86	582
Translate	1.73×10^5	7.01	738
Linear conformal	1.75×10^5	6.95	727
Affine	1.76×10^5	6.97	731

Table 5.1: **Metrics of intensity spread for average images:** It is hoped that well registered images produce a sharp average image, and therefore a wide spread of intensity values. As discussed in the text, none of the metrics were found to be useful.

5.4.2 Training set variance

The aim of registration is to remove variance within the training set that arises as a result of differences in radiographic technique. Therefore, a training set composed of well registered images should have less total variance.
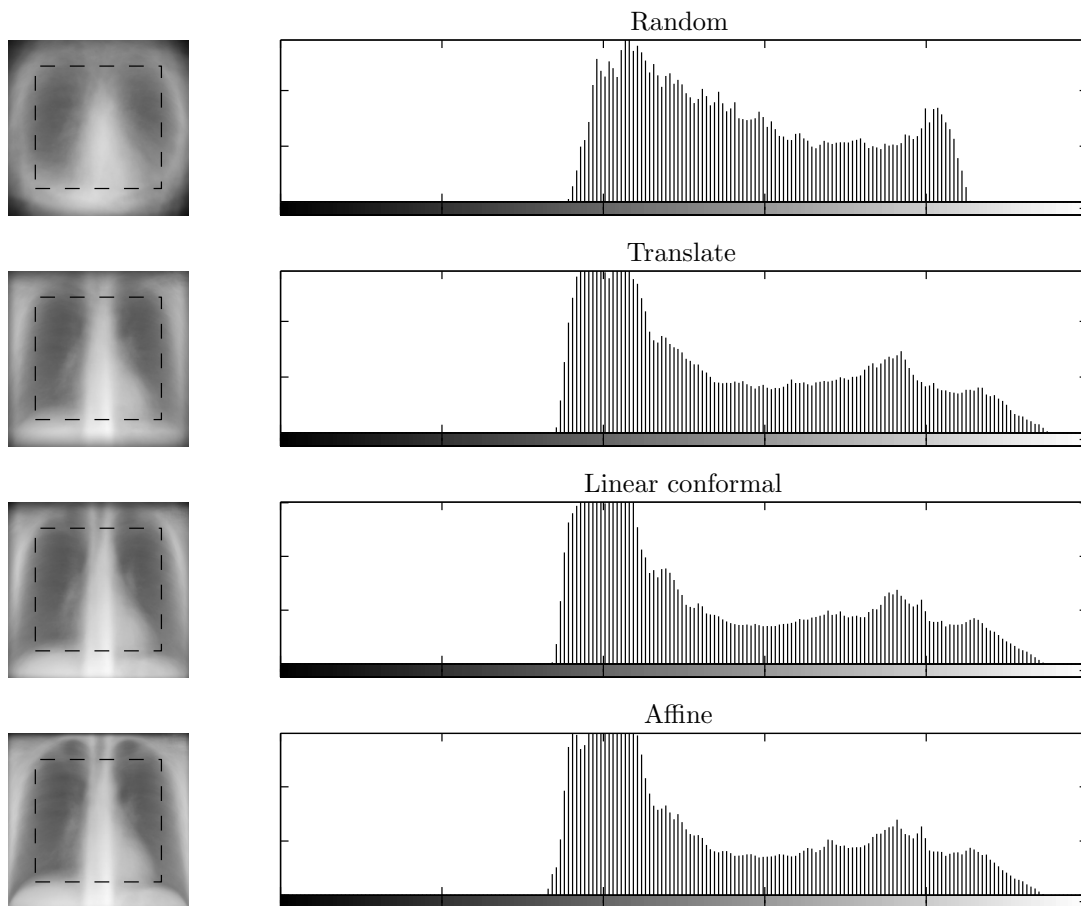
Figure 5.8: **Intensity histograms of average chest images:** The four graphs show the number of pixels at a given intensity within the central portion of each image. The x-axis ranges from black (left) to white (right). Better registration techniques produce average images with more contrast. That is, there are more black and white pixels, and fewer grey pixels. The image variance, IQR, and image entropy aim to quantify this finding.

As introduced in chapter 2, the sum of the eigenvalues of a training set's co-variance matrix is the total variance of the set. Using this knowledge the total variance of the training set was calculated following each of the four semi-automatic methods of registration. The results are given in table 5.2. The results confirms that there is more variance in "worse registered" sets, than the "better registered" sets. Using total variance the order of worse to best registered is: random, translate, linear conformal, affine. This is the order intuitively expected and that apparent by visual inspection of figure 5.6.

Registration Method	Total Variance
Random	7.84×10^7
Translate	5.54×10^7
Linear conformal	3.69×10^7
Affine	3.26×10^7

Table 5.2: **Variance of the training set after four registration methods:** The results confirm that total variance is greater in worse registered set of images that better registered sets of images.

Examining the training set variance does give an indication of how well registered the images are. However, it does not necessarily give an indication of which registration system produces the most useful system. This is because a better registration system will remove more variance, but there is no guarantee that the variance removed during registration is not diagnostic information.

5.4.3 Diagnostic accuracy

The goal of registration is to remove variance that is due to anatomical variation and patient position, but to leave variance which is of diagnostic value. While the methods discussed in sections 5.4.1 and 5.4.2 are interesting and may provide indirect measures of the efficacy of a registration system, the final arbiter of the quality of a registration system is the effect on diagnostic accuracy.

Figure 5.9 plots ROC curves for classification using each of the four registration processes. The best registration method should lead to the ROC curve closest to the top left corner, however the lines for the four ROC curves cross. This implies that for some pathologies one registration routine is better, while

for other pathologies another routine is best. For example chronic obstructive airways disease results in a chest cavity that is hyperinflated. The information regarding chest cavity size would be lost in any registration process that allows the image to be magnified. Thus, no single registration system is ideal for all images. This result was expected as some registration routines will remove variance that contains diagnostic information

The affine routine is the best technique for the widest range of test images, and hence this was used for the remainder of the thesis.
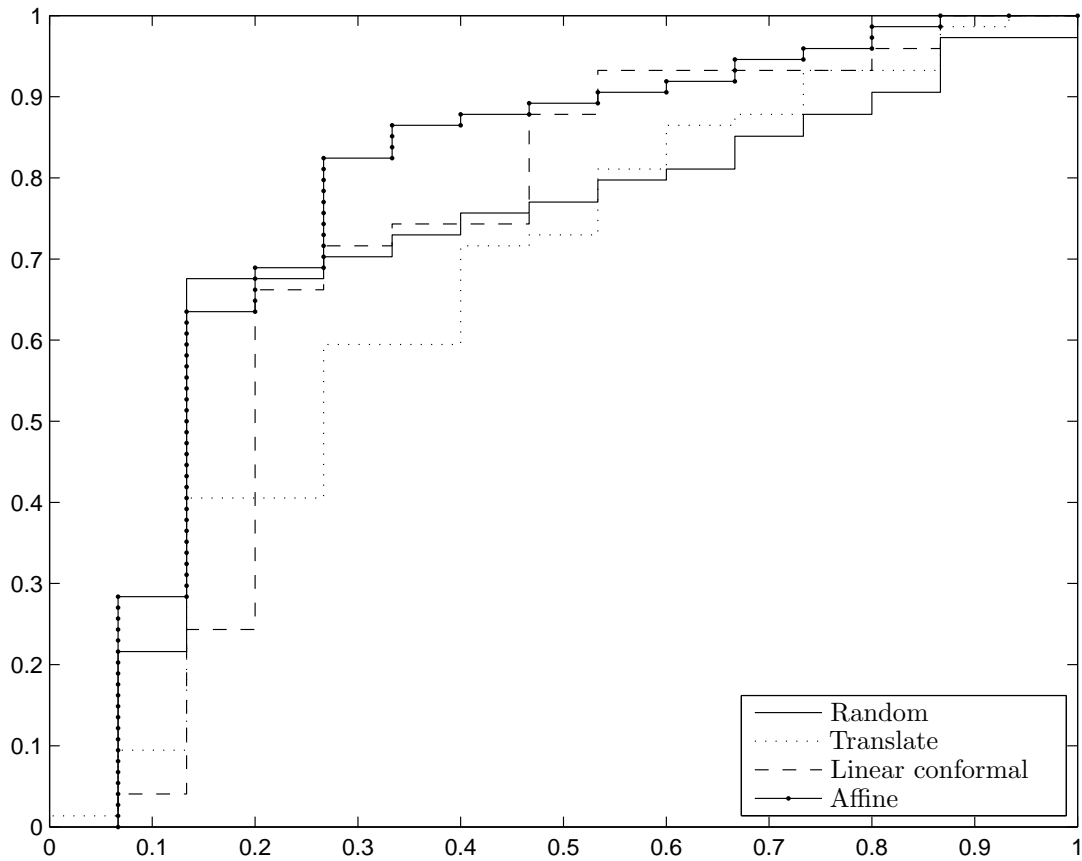

Figure 5.9: **ROC curves from four registration routines:** The graph shows four ROC curves produced using the four different registration systems. The curves cross, suggesting that no one system is ideal, and that while one registration system work best for one image, another system performs best for a different image.

5.5 Conclusion

The aim of image registration is to remove as fully as possible the variation in the test set that can be attributed to the radiographic equipment, methodology, and to the size and build of the subjects. Unfortunately, "non diagnostic variance" is not able to be clearly defined because what is diagnostic information for some diseases is unimportant for other diseases.

While many automated registration techniques have been employed within medical imaging, none are considered ideal for all chest radiograph registration problems. Thus four semi-automatic routines were chosen for evaluation of the effect of registration on the subsequent eigenimage processing. Fortunately, manually identifying control points is acceptable for clinical practice.

Initially, two indirect measures of the quality of registration were assessed: blur of average image, and total variance in the training set. Then a direct measure of the quality of registration was tested: effect on the accuracy of classification.

It was found that while indirect methods provided some insight into how much variance had been removed from the training set, they gave no indication as to whether the registration method removed variance of diagnostic value. Only directly measuring diagnostic accuracy was useful in determining the amount of diagnostic information preserved.

For example, chronic obstructive airways disease is a common chest pathology for which radiographs may lose diagnostic information during registration. In this disease the lungs tend to be bigger than normal subjects, and any registration process that removes the variation in lung sizes will be poor at classifying this disease, despite the images appearing better registered.

Assessment of diagnostic accuracy from the four different registration systems showed that no single registration routine was ideal. Each registration process will have diseases for which the registration process preserves the most diagnostic variance information while removing the most non-diagnostic variation, but no registration process will be ideal for all diseases. There are several possible solutions to this dilemma:

1. Once a complete eigenimaging system is implemented (that incorporates a

particular registration system), test which pathology types are more reliably classified.

2. Use no registration at all, but have a very large training set. This allows the PCA to represent all variance. In this situation the eigenimages would be able to represent all diagnostic information, but at the expense of requiring enough eigenimages to also represent non-diagnostic information.

3. A compromise of the two extremes. For example, use a simple registration system, such as translation only, but have a moderately large training set.

It was found that affine registration generally provided the most predictive value relative to the other methods, although not for all diseases. In any production system that uses eigenimage processing, the registration process would need to be assessed with a view to practical issues (need for user input), image types (chest radiographs versus MRI images), and pathology to be classified.

5.6 Summary

1. Image registration is an important and widely studied topic, but no consensus exists as to what method is most appropriate for chest radiographs.

2. Automatic methods are slow and tend to fail when significant pathology is present.

3. While a registration method may be optimal for one form of pathology, another registration method may be optimal for a different form of pathology.

4. Based on the studies in this chapter, an affine registration system using manually selected control points is a reasonable compromise of simplicity, speed, and reliability.

Chapter VI

Processing of the training set

This chapter discusses the goals, assumptions, and results of the processing which was applied to the training set of normal images in the prototype system presented in chapter 4. Analysis of the processing applied to each test image is presented in chapter 7.

Initially the chapter discusses the goal of Principal Components Analysis (PCA), namely to identify where training images are clustered within the complete image space. Then the dimensionality and spread of the cluster is examined. The effective rank ϵ of A (and therefore of the sample co-variance matrix) is discussed with reference to three common stopping rules used for PCA. The spread of the cluster is analysed with respect to the characteristic roots, λ_i, of the co-variance matrix. A training set containing both pathological and normal images is analysed for dimensionality of clustering.

Following this, there is a discussion of the required number of images in the training set, and to analyse this the effective rank is calculated for a range of different sizes of training sets. Computation speed is measured for the various sizes of training sets used.

Finally, there is a review of some methods for applying PCA to training sets which contain very large numbers of images.

6.1 *Clustering of training images*

The application of PCA to the training set aims to identify a sub-space within the complete image space in which training images are likely to be found. It is expected that images from the training set form a cluster within the complete image space. PCA identifies the centre of the cluster, a set of orthogonal directions which best represent the variance within the cluster, and the magnitude of

variance in each direction.

For the prototype system described in chapter 4, the complete image space is approximately 3×10^6 dimensions, i.e. one dimension per pixel in the images after pre-processing. Eigenimage processing seeks a low dimensional sub-space which spans the majority of the variance within the training set. In an attempt to represent this diagrammatically, figure 6.1 shows a 3-dimensional image space, into which is embedded a 2-dimensional sub-space with basis vectors derived from the training set. In the diagram, image space is $[\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3]$, the training set sub-space is $[\mathbf{u}_1, \mathbf{u}_2]$, and the variance in each dimension of the sub-space is λ_1 and λ_2 respectively. A set of concentric ellipses can therefore be constructed with radii proportional to $\sqrt{\lambda_1}$ and $\sqrt{\lambda_2}$ representing the distribution of training set images. One such ellipse is depicted in figure 2.4.

In order for figure 6.1 to be a useful representation of the formation of a basis set for chest radiographs from a training set certain assumptions must be satisfied. Paraphrasing these assumptions of eigenimage processing in radiological terms gives:

1. There is an average normal radiograph. All other normal radiographs are variations of this.

2. There are a limited number of ways in which a radiograph can vary from the average, yet still be considered normal. For example, one way in which a normal radiograph may vary is heart size.

3. Any variation will have a limited range. For example, the heart can be either slightly bigger or smaller than the average heart, but still be within the normal range.

The first of the above three statements is that an average normal chest radiograph exists. This is represented in figure 6.1 by the location of the centre of the disk with respect to the origin of image space. This assumption is examined in section 6.2.

The second statement is that the training set has substantial variance in only a few directions. In figure 6.1 this is depicted as a 2-dimensional sub-space within the 3-dimensional complete image space. This assumption is tested in section 6.3.
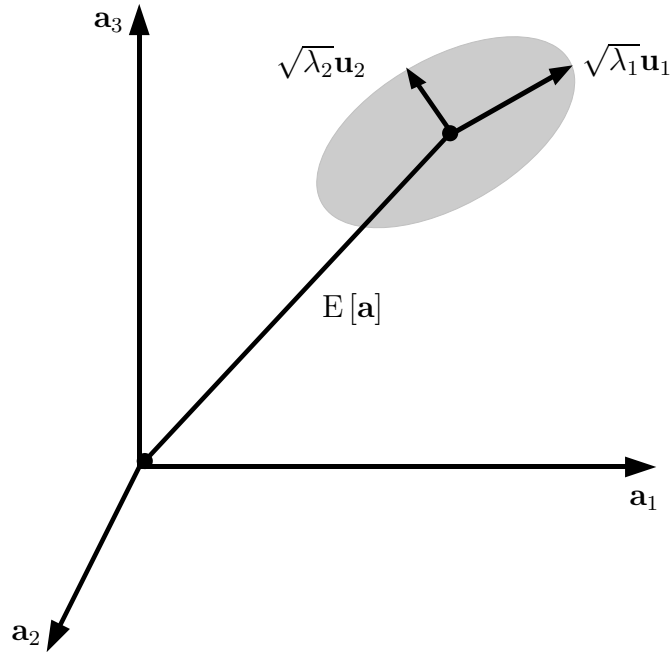
Figure 6.1: **Clustering of training set images:** A depiction of the application of PCA to the training set. In the diagram, the high dimensional complete image space is represented by a 3-dimensional space $[\mathbf{a}_1,\mathbf{a}_2,\mathbf{a}_3]$. The training set images are expected to be clustered. The basis vectors for this sub-space are $\mathbf{u}_1$ and $\mathbf{u}_2$ (principal components). The sub-space is translated from the origin of image space by the empirical mean of the training set, $\mathrm{E}\,[\mathbf{a}]$. The spread of the cluster is related to the variance of the training set along each principal component. Since the variance is given by λ_1 and λ_2 respectively, the training set can be represented as a 2-dimensional elliptical disk with axes of radii $\sqrt{\lambda_1}$ and $\sqrt{\lambda_2}$.

The third statement is that variance in these directions is of limited magnitude. This is represented in figure 6.1 by the sub-space of training images being a disk rather than a plane. This approximation is tested in section 6.4.

6.2   The centre of the cluster

Eigenimage processing assumes that there is an average image about which all other images vary. It is easily calculated by

$$\mathrm{E}[\mathbf{a}] = \frac{1}{p}\sum_{i=1}^{p}\mathbf{a}_i. \tag{6.1}$$

For the training set used in the prototype system the average chest is shown in figure 6.2. All chest radiographs in the training set are variations of the average chest radiograph. However, the utility of finding the average image is minimal unless the other assumptions regarding the training set are shown to be valid.
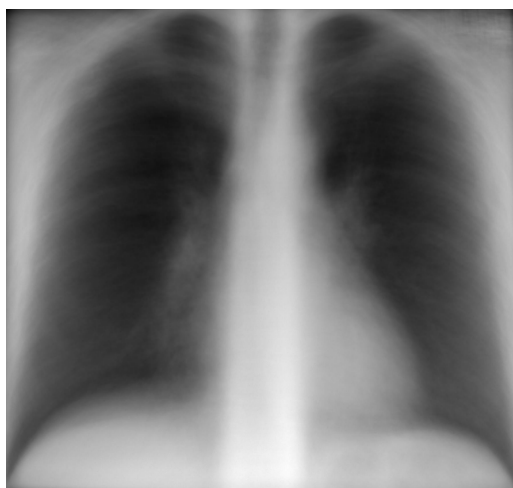
Figure 6.2: **The average chest radiograph:** The image is the average image from the training set of 77 normal chest radiographs that was used in the prototype system.

6.3 The dimensionality of the cluster

This section establishes that there are a limited number of ways in which normal images vary about the average image. Put another way, it establishes the dimensionality of the cluster formed from the training set. Patterns of variance identified by the PCA process will not necessarily be the same as those identified by the human visual system.

In order to determine the dimensionality of the cluster the number of significant components must be determined. The number of significant components is the effective rank of A, denoted ϵ. Characteristic roots for the non-significant components are assumed to be zero within statistical tolerance.

In PCA texts, methods for determining the number of significant components are often referred to as "stopping rules". The goal of stopping rules is to ensure that all retained components represent true signal, rather than noise.

Three methods of determining effective rank are studied in this chapter, although several other methods have been described. The three methods were chosen because: the first is simple and visually intuitive, the second is the most common in software packages, and the third is recommended by major texts [23–25, 60].

Most stopping rules are based on the spectral decomposition of the co-variance matrix, first introduced in section 2.4. The decomposition is represented by

$$\text{var}\,[\mathbf{a}] \;=\; \Sigma \;=\; \frac{1}{p-1}AA^T \;=\; \sum_{i=1}^{r}\lambda_i\mathbf{u}_i\mathbf{u}_i^T \tag{6.2}$$

where p is the number of images in the training set, r is number of principal components and the rank of Σ, while $\mathbf{u}_i\mathbf{u}_i^T$ is one "pattern of co-variance" within the total co-variance matrix. As discussed in section 2.4, $\mathbf{u}_i\mathbf{u}_i^T$ remains "approximately the same size" for all i because $\mathbf{u}_i$ are all unit length. By studying only the relative size of the characteristic roots λ_i (eigenvalues of Σ), it is possible to determine how much of the total variance is represented by only a few principal components.

6.3.1 Scree plot

In 1966 Catell developed the scree plot as a means for graphically determining the relative importance of each principal component of the training set [23, 67]. Figure 6.3 is an annotated version of the scree plot from the prototype system's training set. It demonstrates that there is an "elbow" at around 15 eigenvalues. Using Catell's method, this suggests that only the first 15 components are statistically significant.

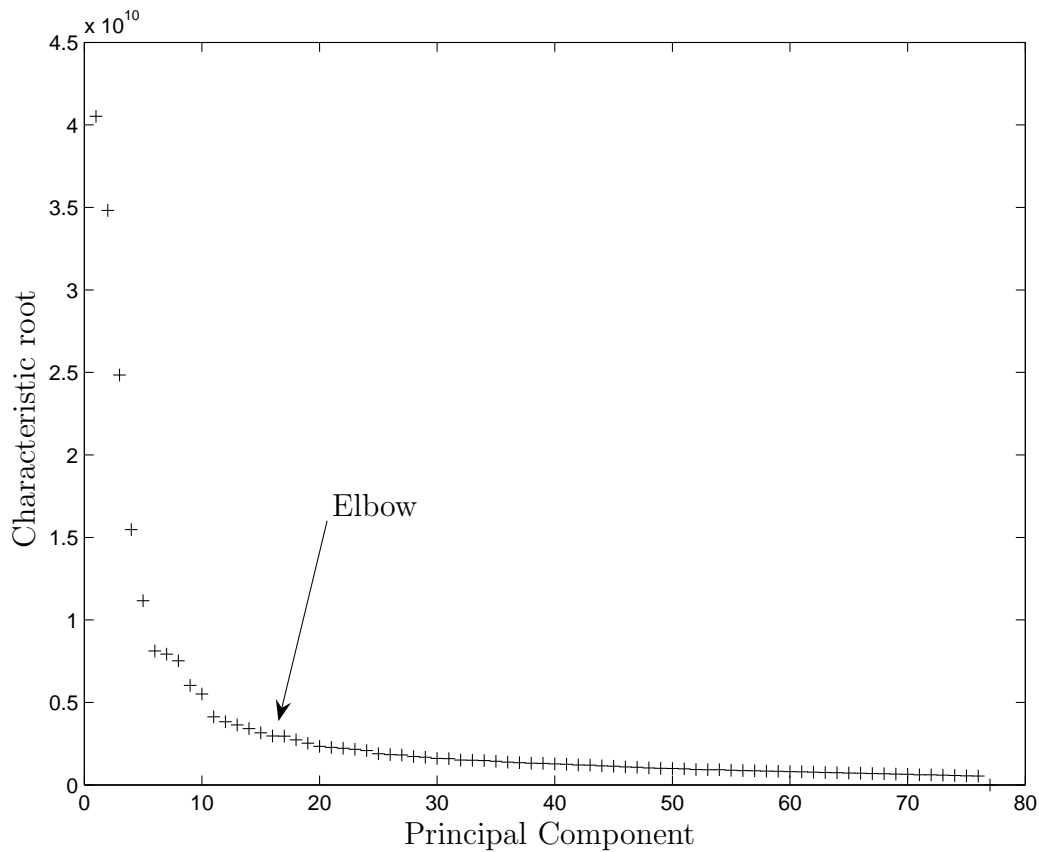Figure 6.3: **Scree plot with elbow:** This demonstrates Catell's stopping rule where the number of significant components is defined by the elbow. For the training set used in chapter 4 the elbow is at around 15 components. This suggests that choosing $\epsilon = 15$ would allow the cluster to span all the significant patterns of variance. This scree plot is the same as in figure 4.5, but is repeated here with the elbow indicated.

The scree plot method is interesting as it provides a visual guide to the number of significant components. However, it is only a visual substitute for a proper significance test and therefore it is not very reproducible. For the examples found in many texts the elbow is easy to identify, but in many real world problems (including chest radiograph analysis) it is much more subjective. For the data presented here it is unclear whether the elbow occurs at 10 components when the data first starts flattening out, or at 20 components when the curve is very flat. Several authors have tried to improve the method, but unfortunately it remains largely subjective.

6.3.2 Power spectrum

The simplest analytical method for determining the effective rank is to choose ϵ so that a known proportion of the total variance is covered. The cumulative proportion of total variation ($CPTV$) is derived from the spectral decomposition of the co-variance matrix given in equation 6.2. $CPTV$ is the proportion of total variance represented by choosing an effective rank of ϵ. Choosing ϵ to include all the components will describe all the variance. $CPTV$ was introduced in section 2.4, and is given by

$$CPTV_\epsilon = \frac{\sum_{i=1}^{\epsilon} \lambda_i}{\sum_{i=1}^{k} \lambda_i} \qquad (6.3)$$

Using data from the prototype system in chapter 4, $CTPV$ is plotted in figure 6.4. The "power spectrum" method allows choosing ϵ to cover a pre-determined proportion of variance. Choosing $\epsilon = 57$ allows normal images to vary in up to 57 independent ways and accounts for 95% of the total variance within the training set of normal images (see figure 6.4).

The method is simple, quick, and found in many software packages. However, many authors have argued that it makes no attempt to determine significant variance. For example if 30% of the variance in the training set is from noise then ϵ should be chosen to cover 70% of $CPTV$. Similarly, if only 1% of the variance in the training set is from noise, then ϵ should be chosen to cover 99% of $CPTV$. The decision of how much variance is due to noise versus signal is left to the investigator, and in practice this decision can be very difficult.
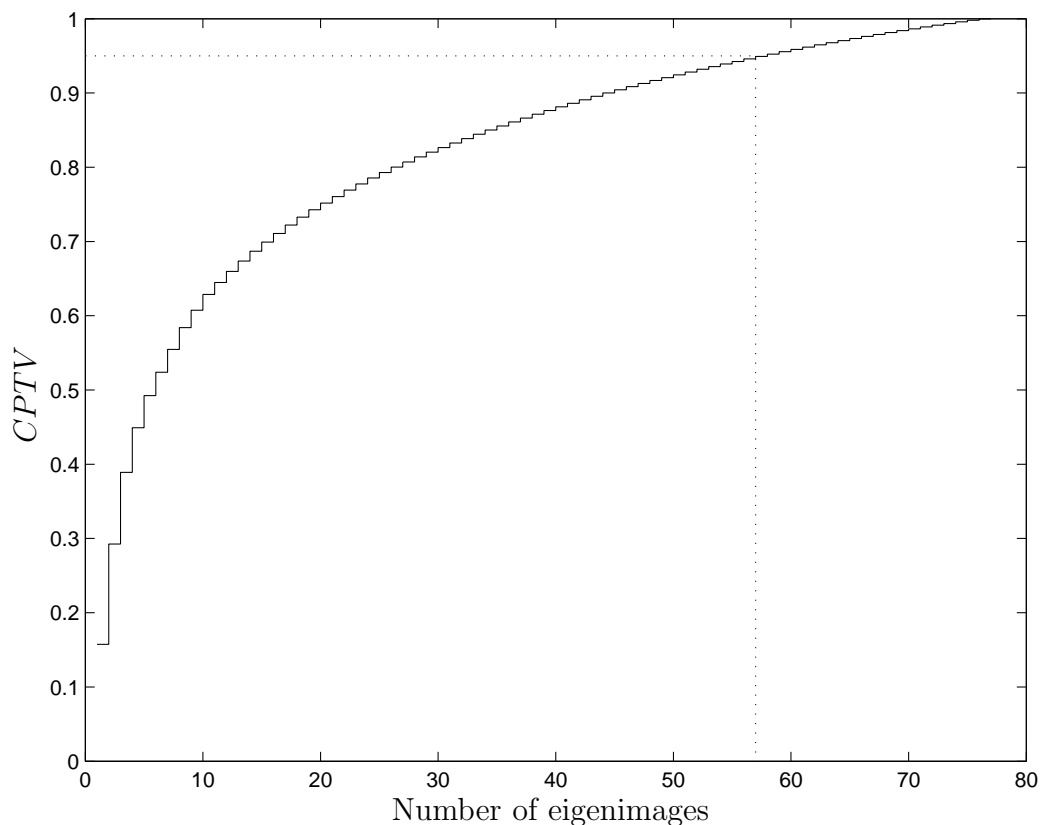
Figure 6.4: **Cumulative proportion of total variance:** Using the training set data from chapter 4 a plot of the proportion of total variance covered by using only a subset of all the eigenimages. The x-axis is the number of eigenimages used, and the y-axis is the proportion of variance represented by retaining a particular number of principal components. Using 57 eigenimages will represent 95% of the variance found within the training set of normal images.

6.3.3   Guttman-Kaiser criterion

A less subjective technique is to retain components with characteristic roots above the average characteristic root. This method is recommended by most major texts, and is sometime referred to as the Guttman-Kaiser criterion [23–25, 60]. It was proposed by Guttman in 1954, but developed further by Kaiser in the 1960's [128, 129].

The rationale is that each component should account for more variance than any single variable in the original set. For example, in a system where each variable in the original data space has a variance of 1, significant components should all have a variance greater than 1. Using this rationale it can be shown that for unscaled data the cut-off for significant characteristic roots is the the average characteristic root $\bar{\lambda}$. Details can be found in most PCA texts [23–25, 60].

Since this method was first proposed, there have been a significant number of publications attempting to fine-tune it. In the 1970's Jolliffe studied the method using both simulated and real data and concluded that the cut-off should be approximately 70% of the average characteristic root [130, 131].

Applying the Guttman-Kaiser criterion with $\lambda_i \geq \bar{\lambda}$ to the data used for the prototype system in chapter 4, it is found that the training set has an effective rank $\epsilon = 14$. This equates to 14 significant principal components. Using a cut-off of 70% of the mean characteristic root, $\lambda_i \geq 0.7\bar{\lambda}$, suggests that an appropriate effective rank is $\epsilon = 19$.

6.3.4   Effective rank with pathology included in the training set

For the prototype system the training set only included normal images. If the goal is to have eigenimages that can closely represent both normal and abnormal images it is likely that the training set will need to include pathological images. The benefits of having eigenimages that can represent pathology are discussed in section 7.5. However at this point it is worth considering the effect of pathology on clustering of the training set.

It is expected that adding pathological images will increase the dimensionality of the cluster. That is, by including pathology new patterns of variance will be introduced into the set. However, unless a sufficient number of examples of a particular pathology are introduced, the extra patterns of variance may be of
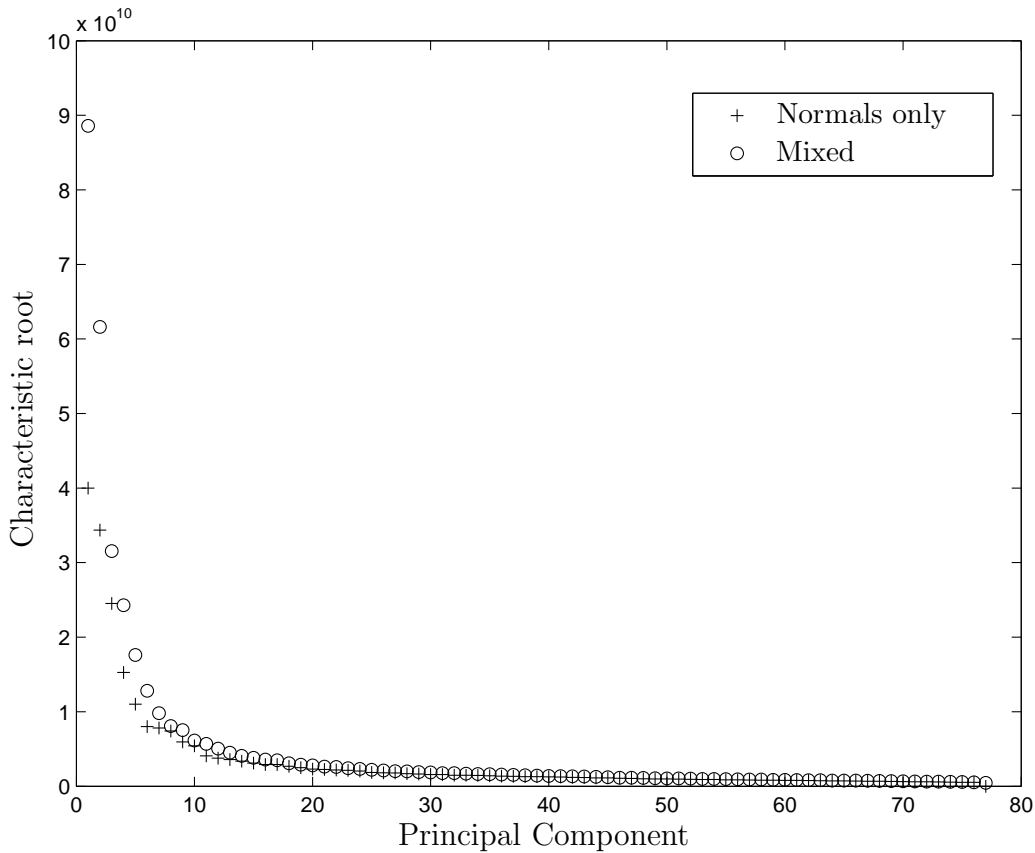
Figure 6.5: **Scree Plot of training sets with and without pathology:** The two plots show the characteristic roots for training sets of containing 77 images. The first set contains only normal images, while the second set contained a mix of 38 normal and 39 pathological images.

such a low power that they are indistinguishable from noise. For example, if the training set includes only one image of a right upper lobe pneumonia, insufficient variance may be added to produce a significant eigenimage.

To test the effect of including pathology, PCA was applied to a new training set containing a mix of 38 normal images and 39 abnormal images. The total size of the new mixed training set was the same size as that used in the prototype system. Pathological images included pneumonia, lung masses, and heart failure.

Scree plots of both the original normal training set and the new mixed training set are shown figure 6.5. The figure demonstrates that the characteristic roots

of the training set containing a mix of pathological and normal images tend to be larger than the training set containing only normal images, implying there is more variance within the mixed set.

Using the Guttman-Kaiser criterion at a significance level of $0.7\bar{\lambda}$ the effective rank of the training set containing pathology is 17, which is less than the effective rank of the training set containing only normal images. The reduced effective rank in the mixed training set suggests that while the pathological images introduce more patterns of variance, these patterns have relatively low power. This is because there is a wide variation in pathology. For example, two images of pneumonia may be very different: one may have consolidation in the right lung while the other has consolidation in the left lung. Introducing these images into the training set introduces two new patterns of variance, both of low power. If more examples of pneumonia were added then the power of these new patterns of variance would increase, thus making them more likely to be significant.

In summary, it is possible to include pathology in the training set of images. Unfortunately, due to the wide variation of appearances of "similar pathology", it will be necessary to have many examples of each form of pathology.

6.3.5 Choosing the effective rank of the training set

The three most popular methods for choosing the effective rank of the training set, reviewed above, have indicated a range of possible values for the prototype system. The range of values is $\epsilon = 14 \ldots 57$. This range reflects the uncertainty in the statistical literature concerning which method is most appropriate.

Neither the scree plot nor the power spectrum method are ideal for eigen-imaging. While the scree plot is conceptually useful it is remains a subjective test. The power spectrum is good at determining the relative variance of each component, but to be useful requires an estimate of the expected signal-to-noise ratio. i.e., if the noise is expected to be 5% of the total variance, then the effective rank ϵ should be chosen to span 95% of the total variance. For sets of images, such as in the prototype system, it is difficult to estimate the signal-to-noise ratio.

The Guttman-Kaiser criterion is more sophisticated because it estimates which components are significant based on a measure intrinsic to the data. This

method is still not ideal because there remains doubt as to exactly where the cut-off for significance should be. However, it is an improvement over the power spectrum method. Using the Guttman-Kaiser criterion with a 70% cut-off, the effective rank is $\epsilon = 19$.

In practice, for many eigenimaging problems the chosen effective rank is relatively unimportant since the goal is to identify the important patterns of variance in order to describe of a cluster of normal images. For many problems, identifying the important (common) patterns of variance is adequate, i.e. low order components. Remaining patterns of variance (i.e. high order components) may or may not be statistically significant, but certainly the variance they represent is of less magnitude than lower order components.

6.4 The spread of the cluster

Having identified the principal directions in multi-dimensional space along which normal images vary about the mean, it is now possible to consider the magnitude of their variance in each of these directions. With reference to figure 6.1, this is the spread of images along each principal component.

From material on PCA presented in chapter 2, it is known that for any particular principal component $\mathbf{u}_i$, the variance along the component is given by the associated characteristic root λ_i. As mentioned in 3.6.1 the characteristic roots of the sample co-variance matrix $\Sigma = \frac{1}{p-1}AA^T$ can be found from the singular values of A by $\lambda_i = \frac{s_i^2}{p-1}$. At this stage no reference has been made to the shape of the distribution along each component, merely the variance of the distribution.

Many theorems of PCA stopping rules assume that the original variables have a multivariate normal distribution and therefore assume that the distribution with respect to each principal axis is also normal. One of the few papers to study the distribution of chest radiographs in image space concluded that a multivariate normal (Gaussian) distribution is satisfactory [52].

Waternaux and Davis showed that even if the original variables are not multivariate normal, the distribution along each principal axis is still expected to be normal so long as Σ and the fourth cumulant of the underlying univariate distributions are finite [132, 133]. The fourth cumulant (called kurtosis) is given

by

$$\text{kurt}[x] = \text{E}[x^4] - 3(\text{E}[x^2])^2 \qquad (6.4)$$

where x is a single original variable, i.e. a single pixel's intensity distribution in the source images. Since each source pixel's intensity is finite and small, $\text{E}[x^4]$ and $\text{E}[x^2]$ should be also finite. Thus the kurtosis should be finite, again leading to the expectation that the distribution along each principal component is Gaussian, with variance λ_i.

Weighting factors for the images in the training set from the prototype system in chapter 4 were calculated as per section 4.6.3. Each weighting factor gives the location of a training image along each associated principal component. Four example distributions of weighting factors are shown in figure 6.6, along with superimposed Gaussian distributions each with a variance of λ_i. As expected, the spread of images decreases with higher order principal components. This is expected because the characteristic roots λ_i are smaller for higher order components.

Figure 6.6 visually compares the distributions to the expected Gaussian distribution. To test this expectation, the Kolmogorov-Smirnov test can be used. This is a widely accepted statistical test to establish whether a sampled distribution could be from a known distribution [134]. The Kolmogorov-Smirnov test analyses the data to accept to reject the null hypothesis that the distribution of the training set images' weighting factors along the ith principal component is a Gaussian distribution with mean $= 0$ and variance $= \lambda_i$.

The two-sided Kolmogorov-Smirnov test computes a test statistic which is the maximum difference between the sampled data and a specified probability distribution, and if the test statistic is above a critical value the null hypothesis is rejected. The critical value is a function of the confidence interval and sample size. For these tests the confidence interval was chosen to be 99%, that is a p-value=0.01, with a sample size of 77. The test is said to be two sided because it allows the sampled distribution to be either to the left or right of the predicted Gaussian distribution.

The results of the Kolmogorov-Smirnov test are shown in figure 6.7. No test statistic is greater than the critical value and hence the null hypothesis cannot be rejected for any principal component. Specifically, this means that based on
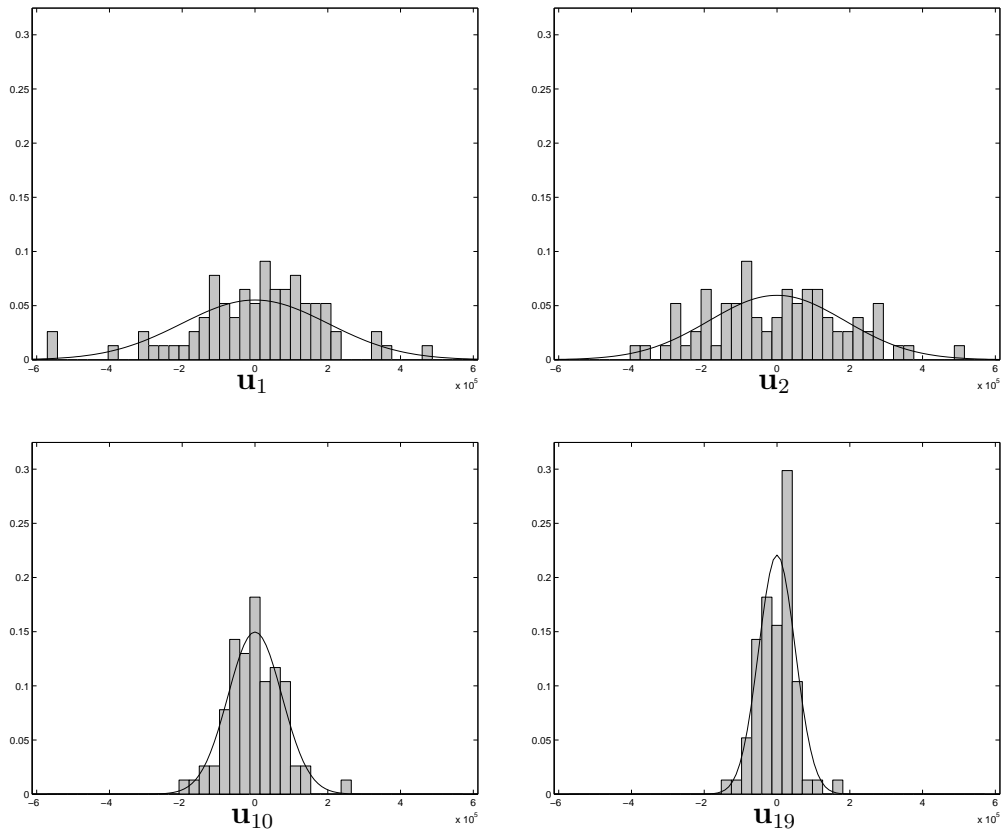
Figure 6.6: **Spread of the training set cluster:** The four histograms give the spread of the weighting factors from training set images. Four example of principal components are shown: $\mathbf{u}_1$, $\mathbf{u}_2$, $\mathbf{u}_{10}$, and $\mathbf{u}_{19}$. Each bar of a histogram represents the proportion of images with a weighting factor in a particular range. Plotted on top of each histogram is a Gaussian distribution of variance λ_i, normalised so that the area under both the histogram and Gaussian distribution are equal.
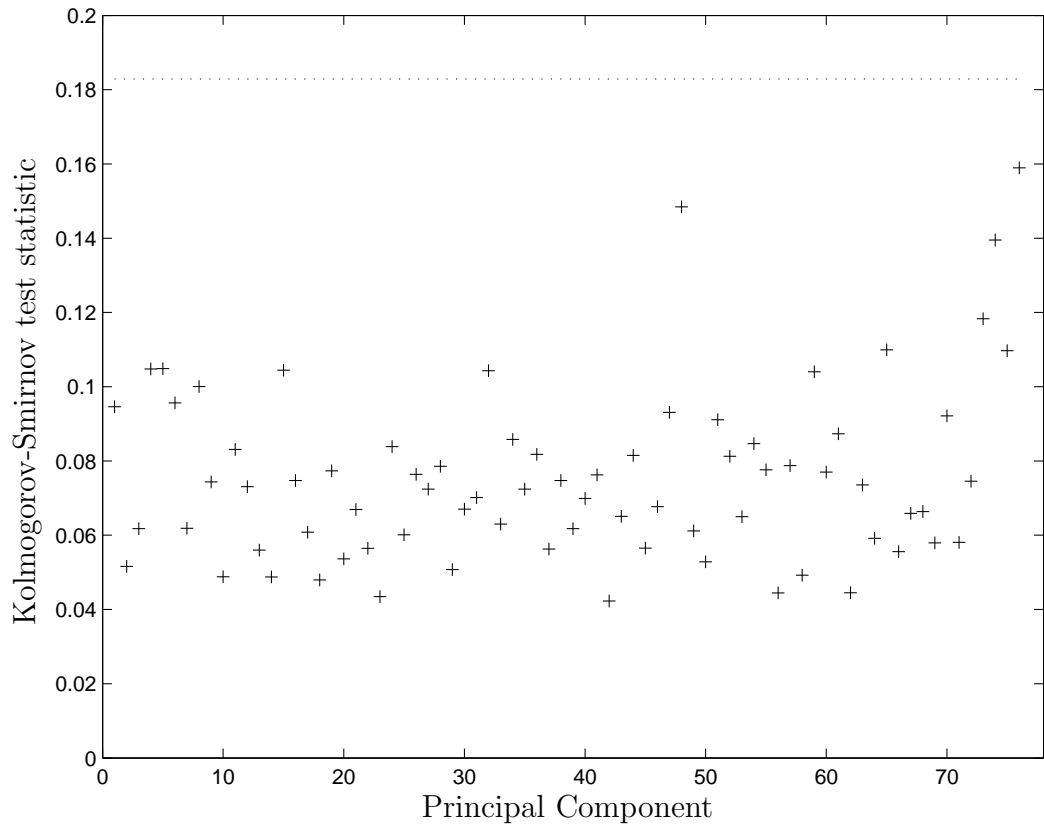
Figure 6.7: **Kolmogorov-Smirnov test:** This plot shows the Kolmogorov-Smirnov test statistic for each principal component. The dotted line is the critical value required to reject the null hypothesis for a p-value of 0.01. At this confidence level the null hypothesis was unable to be rejected for any principal component.

the sample, the distribution of weighting factors for each principal component is indistinguishable from a Gaussian distribution.

6.5 Estimating the number of training images required

In the prototype system the size of the training set was based on set sizes used for face recognition systems. It is important to develop methods to estimate the required training set size for chest radiographs, as this will influence how useful eigenimage processing is for radiological images.

For face recognition many methods have been used to estimate the required

training set size. The metric used to determine the number of required images depends on the specific use of eigenimaging. For example, when designing a face recognition system the training set may need to describe all possible human faces to within a certain prescribed error. In addition, the number of images required varies depending on the quality of images within a training set. For example, variation in image size, lighting conditions, features (such as sun glasses), all affect the required number of training images. For face recognition, authors have claimed that between 20 and 200 eigenfaces are needed to cover an "acceptable" range of variation [135, 136].

The chest radiograph system aims to have the principal components represent a reasonable amount of normal variation. Unfortunately, for chest radiographs there is an extremely large number of identified "normal variants", some of which occur at a frequency of less than one in tens of thousands of people [85, 137, 138]. Thus it will never be possible to have a set of principal components that is able to represent all normal variants. However it will be useful to develop some understanding of the effect of changing the number of training images.

6.5.1   Experimental design

A series of experiments was performed to clarify the effect of changing the number of training images. The goal was to find the characteristic roots, total variance, and effective rank for a range of training set sizes. For these experiments the total set of images included the 77 normal training images as well as the 15 normal test images, thus giving a maximum set size of 92 images. This was achieved by starting with a small training set, $p = 2$, then successively adding images until the training set contained all the normal images. All images had previously been registered as per the prototype system.

For all sizes of training set, $p = 2 \ldots 92$, the following calculations were performed:

1. The training set was constructed $A = [\mathbf{a}_1 \ldots \mathbf{a}_p]$.

2. The average image $\mathrm{E}[\mathbf{a}]$ was found and subtracted from all the images.

3. $A^T A$ was found and its eigenvalues (s_i^2) calculated.

4. The characteristic roots of Σ were derived and stored using the relationship $\lambda_i = \frac{s_i^2}{p-1}$.

5. The time taken to perform these calculations was recorded.

6. Total variance and effective rank were calculated from the characteristic roots. Both the Guttman-Kaiser criterion and the power spectrum method were investigated.

The order of the images was then randomly shuffled, and the above process repeated for a total of $n = 30$ runs. Using this method it is worth noting that for small training sets, e.g. $p = 20$, it is likely that the composition of the training set for one run is very different from the other 29 runs. Conversely for large training sets, e.g. $p = 90$, it is likely the composition of the set will be very similar to the other 29 runs performed with $p = 90$.

Overall, the experiment involves finding characteristic roots for $n \times p_{max} = 2730$ training sets, with the training set sizes ranging from $p = 2 \ldots 92$ images. The total time to run the calculations was approximately 200 hours.

6.5.2 Changes in total variance with training set size

As the number of images in the training set increases it is expected that the total variance will increase until the sample co-variance matrix is formed by enough samples to closely approximate the population co-variance. As discussed in section 2.4, the total variance (TV) in a training set can be found by summing up all the characteristic roots. For the mth run this was done using the characteristic roots calculated for training set.

$$TV_m(p) \quad = \quad \sum_{i=1}^{p} \lambda_{i,p,m} \tag{6.5}$$

The total variance was then averaged over the $n = 30$ runs. The average $TV(p)$ is plotted in figure 6.8 with plus and minus one standard deviation plotted as dotted lines.

When examining the data in figure 6.8 the mean of the 30 runs for calculating $TV(p)$ remains approximately constant for increasing p. This is expected because
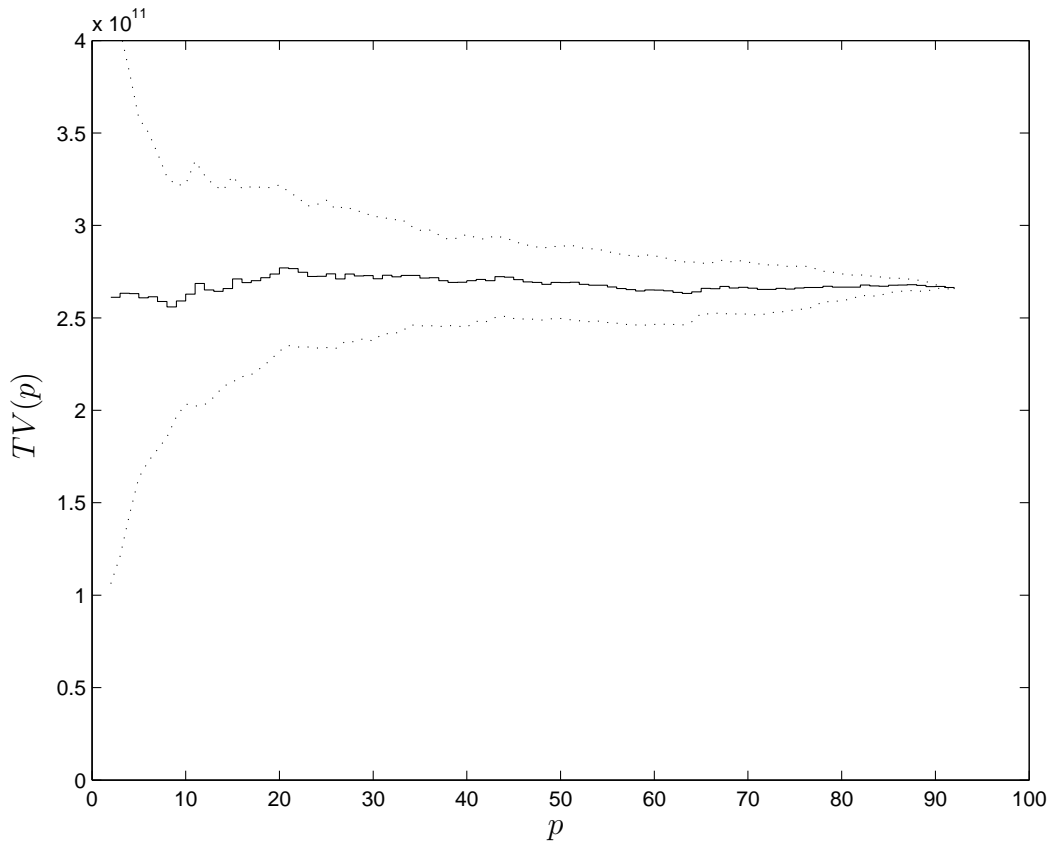
Figure 6.8: **Changes in total variance with set size:** The plot demonstrates that the total variance (sum of the characteristic roots) of the sample co-variance matrix tends to remain constant with increasing the number of images in the training set. The sold line is the average over the $n = 30$ runs for $TV_m(p)$, while the dotted lines are plus and minus one standard deviation.

a sample co-variance matrix should be an estimate of the population's co-variance matrix and thus remain approximately constant over a range of sample sizes.

Also in figure 6.8 the range of $TV_m(p)$ over the $n = 30$ runs (as depicted by the dotted one standard deviation lines) tends to decrease with increasing p. This is expected for two reasons. Firstly, as the sample size increases the sample co-variance matrix becomes a better estimate of the population co-variance matrix, which leads a smaller range of $TV_m(p)$. Secondly, as discussed previously, for large values of p the training sets are almost identical and therefore the range of

$TV_m(p)$ will be small. Unfortunately, it is difficult to estimate which of the two factors is most important.

6.5.3 Changes in effective rank with training set size

As the number of images in the training set increases it is expected that the number of significant components will increase and therefore that the effective rank of A and Σ will increase. This is because introducing new images can either introduce new patterns of variance or make patterns already present statistically significant.

For example, the training set may contain only a few examples of a normal anatomic variant called an "azygous fissure". This variant is normally present in around 2% of chest radiographs, but is a small low intensity finding. Consider that the variant is present within the training set, but that it is not statistically significant using a test such as the Guttman-Kaiser criterion. Introducing a new image containing this variant into the training set may allow the component to become statistically significant, thus increasing the effective rank of the training set. Then there is the question of which normal variants should be included in the normal training set, as some normal variants are clinically important while others become clinically important in conjunction with abnormal findings.

If the training set is small, when a new image is introduced it is likely that the effective rank will increase. This is because it is likely that the new image contains either a new pattern of variance or a pattern of variance that was previously not statistically significant. If the training set is large, introducing a new image is less likely to increase the effective rank. This is because if a training set already well represents the population, it is unlikely that introducing a new image will introduce a new pattern of variance, because such a pattern is likely to already be represented. In summary, if the training set is large enough and represents the population well, the rate of change of the effective rank should be slow when increasing the number of images in the training set.

Section 6.3 described two methods by which the effective rank of Σ can be automatically calculated: the power spectrum method and the Guttman-Kaiser criterion. While the scree plot also finds the effective rank, it requires manually identifying the "elbow" and is therefore subjective and less easily automated. To

Figure 6.9: **The power spectrum method for determining effective rank as a function of the number of training images:** The effective rank was calculated to span 50%, 70% and 90% of the total variance, plotted as ϵ_{50}, ϵ_{70}, and ϵ_{90} respectively.

test the effect of changing the size of the training set has on effective rank, ϵ was calculated for a range of sizes of training sets, i.e. for a range of values of p. Both the power spectrum method and Guttman-Kaiser criterion were used to determine the effective rank.

Using the power spectrum method to determine effective rank requires an estimate of the relative signal-to-noise ratio for the system. The value of ϵ is chosen to span the estimated significant fraction of the total variance. The effective rank was determined so that the principal components would span 50%, 70% and 90% of the total variance. The effective rank as a function of set size (ϵ_{50}, ϵ_{70}, ϵ_{90}) was then averaged over the $n = 30$ runs. This is plotted in figure 6.9.

It can be seen that when choosing to cover a larger portion of the total variance, e.g. ϵ_{90}, the effective rank tends to be larger and rapidly increases with an increasing number of images in the training set. This is thought to be because adding extra images to the training set adds more low power variance (which is spanned by the low order eigenimages). However if a smaller portion of total variance is required to be covered, e.g. 50%, then the effective rank tends to be relatively stable after approximately 50–60 training images. This stability is thought to be because adding an extra image is unlikely to alter high order eigenimages (which represent the strongest 50% of variance).

Based on this experiment it is believed that to span any given proportion of variance, there is a training set size for which the effective rank is relatively stable with increasing training set size. Choosing to span 50% of the variance in chest radiographs requires that training set should contain at least 50–60 images for the effective rank to become stable. Similarly, choosing to span 70% of the variance requires a training set of 80–100 images for the effective rank to be stable, and choosing to span 90% of the variance requires a test set containing more images than used in this study. It is postulated that a set size of 200–300 images may allow the effective rank to be stable with changes in set size, although this cannot be tested using available data.

As previously introduced in section 6.3.3, the Guttman-Kaiser criterion considers principal components significant if the associated characteristic root λ_i is greater than the average characteristic root $\bar{\lambda}$. This method can be fine-tuned by adjusting the significance level relative to the average characteristic root. A widely accepted value for significance is $\lambda_i \geq 0.7\bar{\lambda}$ [130, 131].

Figure 6.10 shows how the effective rank, determined using the Guttman-Kaiser criterion, changes with increasing set size. The results imply that choosing components that have characteristic roots $\lambda_i \geq 1.3\bar{\lambda}$ to be significant, requires that the training set should contain 90-120 images for the effective rank to be stable. However, using the preferred cut-off of $\lambda_i \geq 0.7\bar{\lambda}$ and extrapolating from figure 6.10 it is again postulated that the effective rank will not be stable until at least 200–300 test images are used.

Comparing the power spectrum method with the Guttman-Kaiser criterion, a similar stability pattern is found for the effective rank as a function of the number of training images. That is, when the number of training images is large
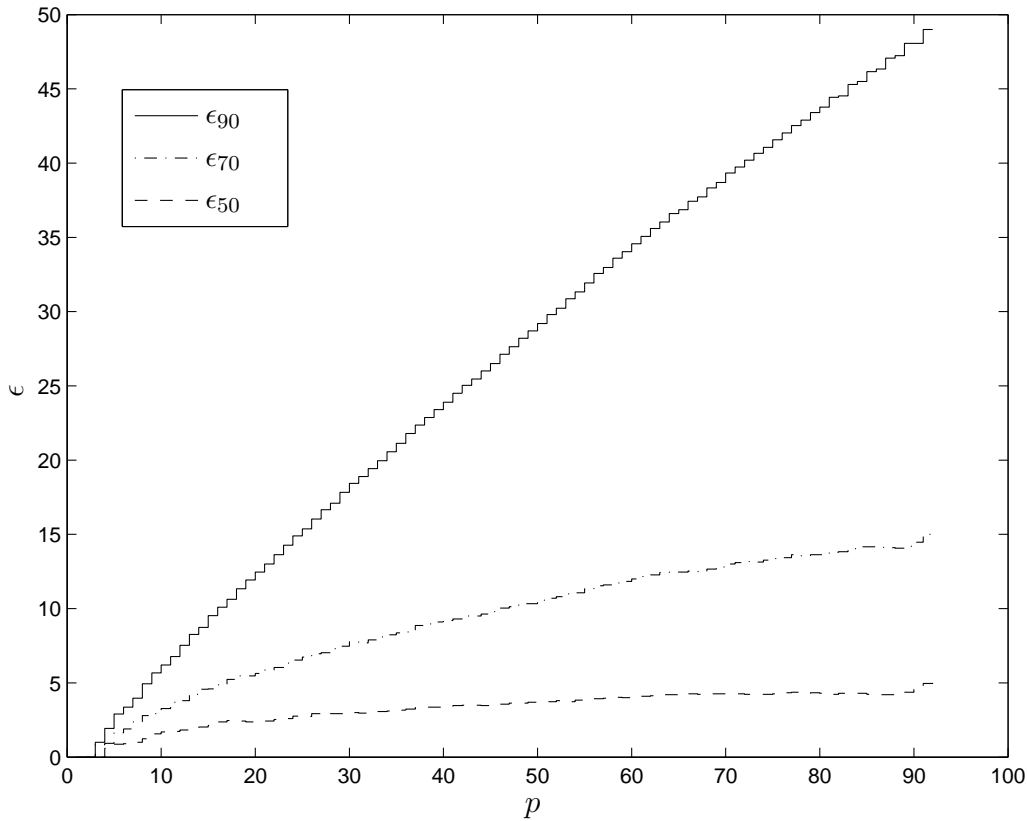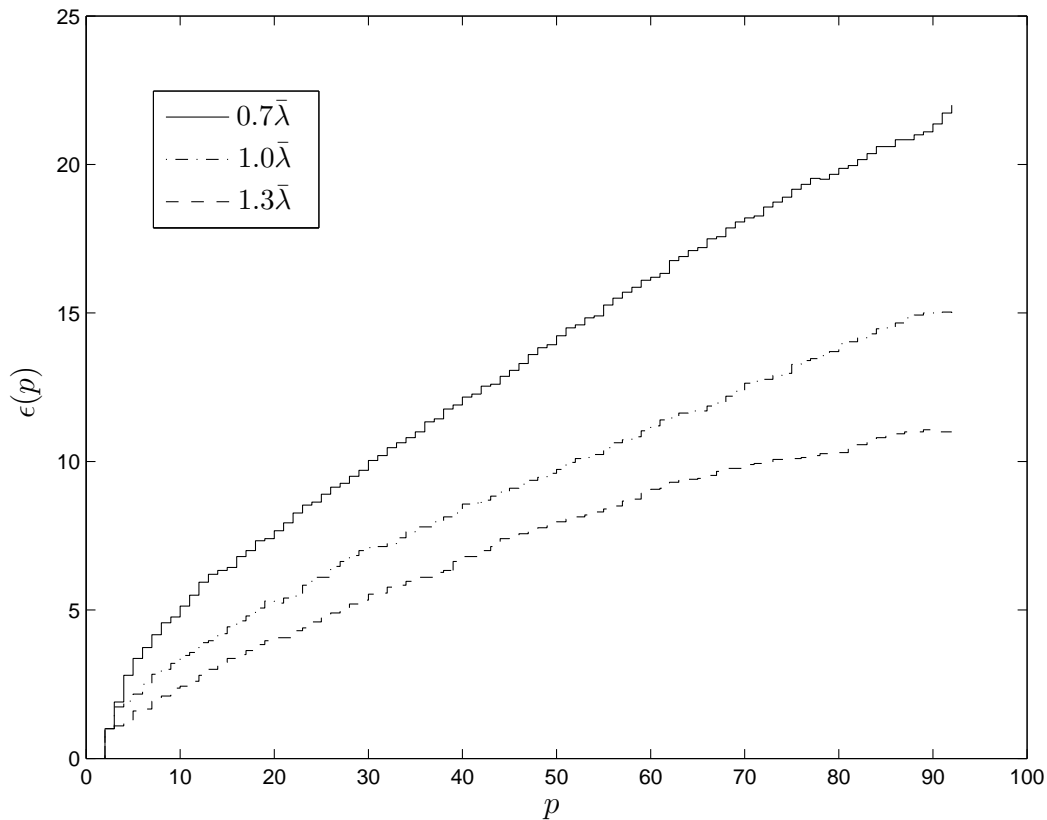
Figure 6.10: **The Guttman-Kaiser criterion for determining effective rank as a function of the number of training images:** The effective rank is determined by totalling the number of characteristic roots above the average. Early work by Guttman and Kaiser suggested a suitable cut-off was $\lambda_i \geq \bar{\lambda}$. Later work by Jolliffe suggested that $\lambda_i \geq 0.7\bar{\lambda}$ is more appropriate. For comparison $\lambda_i \geq 1.3\bar{\lambda}$ has also been plotted.

the effective rank tends to stabilise (figure 6.10). However, the number of images in the training set required for stability depends on the exact cut-off relative to $\bar{\lambda}$.

In conclusion, using the Guttman-Kaiser criterion with $\lambda_i \geq 0.7\bar{\lambda}$, the effective rank of the training set in the prototype system (77 images) is $\epsilon = 19$. It is expected that adding more images to the training set would alter this effective rank. If the training set was greater than 200–300 images the effective rank should be less sensitive to changes in training set size. Unfortunately, due to difficulty in obtaining enough images, sets of this size could not be evaluated for this study.

6.6 Computation speed

The tests in section 6.5.3 required performing eigenimage processing 30 times on sets with $p = 2\ldots 92$ images. This gave an opportunity to time the processing for a variety of training sets, thus allowing evaluation of speed. When this research was started in 2002, processing for the prototype system, $p = 77$, took around 15–30 minutes. By May 2006, there had been significant improvements in computing hardware and the processing time was around 2–3 minutes. Although computations were done using a small Linux cluster, the code developed did not take advantage of parallel processing and therefore only ever ran on one node of the cluster. These were 1 GHz Pentium-III processors with 512 Mb of RAM.

As introduced in section 4.4, the time for eigenimage processing of the training set increases linearly for increasing k and quadratically for increasing p. Figure 6.11 plots the average time taken for 30 runs of calculating $A^T A$ and its eigenvalues. This is the computationally intensive part of eigenimage processing. Using the graph to extrapolate to training sets of 200–300 images, it is expected the processing time would be around 10 minutes. An experiment was performed in which the training set images were duplicated four times, giving a total set size of 308 images, and processing for this very large set took 12 minutes.

The code used to find $A^T A$ was introduced in figure 4.2. This code has two instructions involving a significant computation time. The first is reading in a row of A from disk. The total time for this should be a linear function of k (the image size) and p (the number of training images). The second slow instruction is finding C from the row of A. The total time for this instruction should be
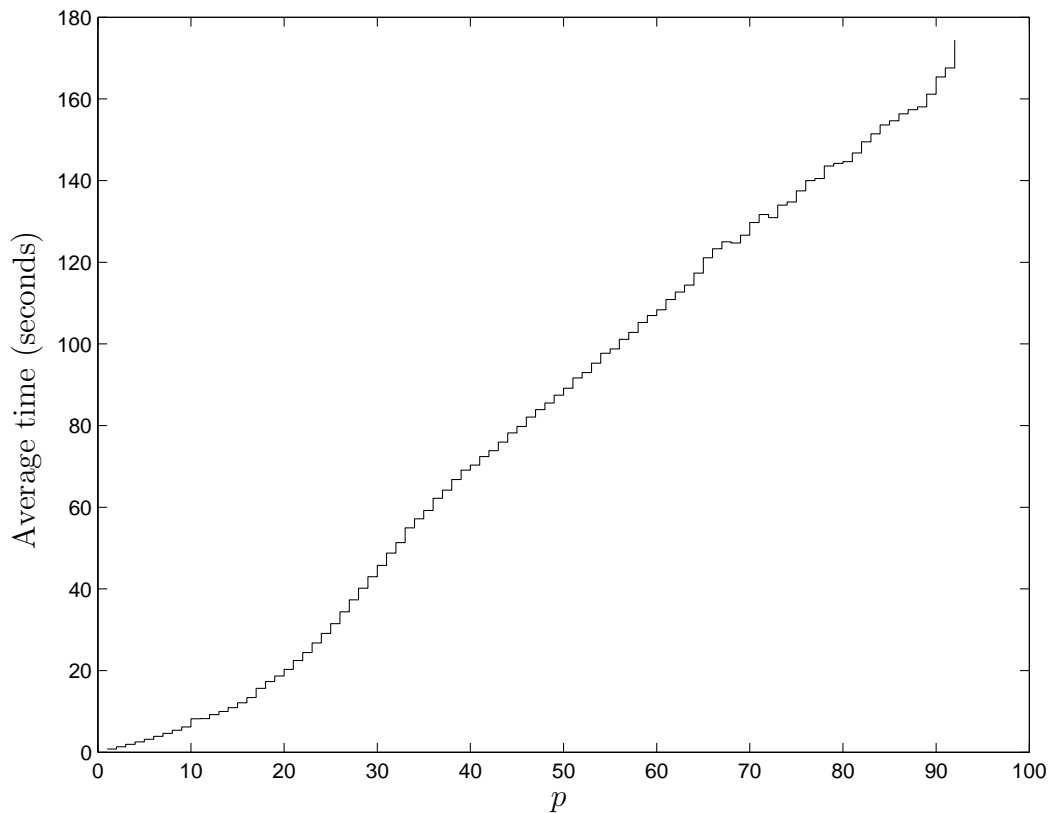
Figure 6.11: **Time for processing of the training set:** Plot of the average time to calculate $A^T A$ and it eigenvalues, for $p = 2 \dots 92$ images.

a quadratic function of p. Looking at figure 6.11, the graph is mostly linear suggesting that for training sets up to $p = 92$ images the time to read all of the rows of A from disk is more significant than the time to do the calculation for finding C.

The finding that calculating A^T is limited by disk access speed rather than CPU speed was expected. This is because for each iteration of the loop in figure 4.2 one row of A is read in, then C is computed. For $p = 92$, reading in one row requires reading from 92 separate files stored on a network disk, which is relatively slow. Computing C requires 92^2 multiplications and additions, which should be very fast. Even if the disk access was fully optimised it is likely that reading in the data will still remain much slower than the calculations.

6.7 Altering the training set

Using the method described for the prototype system it should be possible to calculate principal components for sets of radiographs containing 200–300 images. This should be an adequate set size for the applications described in this thesis, but there are occasions when it may be helpful to have either a very large training set, or to regularly update the training set. There are several ways to do this. Different names have been given to these methods, but incremental PCA and sub-space PCA are commonly used terms [29, 45].

There have been many methods to incrementally adjust the size of the training set, and these forms of PCA are commonly known as incremental PCA. However, most methods have significant drawbacks. For example, while some methods allow for updating of the average image and the eigenimages, they are unable to update the eigenvalues [139, 140]. Other methods update both the eigenimages and eigenvalues, but assume that the average image remains constant [141]. One method updates the eigenimages, eigenvalues, and average image but requires keeping track of the full co-variance matrix, which is not feasible for large images [142].

Another method for performing PCA is to divide the training set into parts. PCA is then applied to each part, thus finding multiple sub-spaces which are then added. These techniques are often called sub-space PCA. A general method of sub-space PCA is reviewed in section 6.7.2.

6.7.1 Motivations for altering the training set

There are several situations where it may be desirable to use a different approach for performing PCA calculations than the approach presented in this thesis. Possible motivations for adopting one of these more flexible forms of PCA include:

Large training sets: A large training set may be required for radiological images because of the large number of normal variants encountered. To construct these large training sets, smaller sets may be added together [143].

Update the training set over time: As time passes it may be desirable to update the set. This may be because of changes in the sample population or variation in imaging techniques.

Including pathology in the training set: A training set could be formed using sub-spaces such as normal, pneumonia, heart failure, or other pathology. Each sub-space needs only enough images to represent that pathology well, e.g. 300 normal images, 500 images of pneumonia etc. There is no requirement for the relative population frequency of each pathology to be known.

6.7.2 A general method for altering the training set

An important method of sub-space PCA was introduced by Hall *et al.* [144–146]. Their method allows for adding and subtracting eigenspaces, while keeping track of the eigenimages, eigenvalues, and average images without the need to compute the entire co-variance matrix. The method can be used to add or subtract either one or many images from a training set, thus performing either incremental PCA or sub-space PCA.

The method involves considering the Singular Value Decomposition (SVD) of a matrix X

$$A_X - \bar{\mathbf{a}}_X \mathbf{1} = U_X S_X V_X^T \tag{6.6}$$

where A is of k rows and p_X columns and $\bar{\mathbf{a}}_X \mathbf{1}$ is the empirical mean of the set. The eigenspace $\Theta(X)$ is defined to correspond to the first p_X columns of U_X, V_X, and S_X. Hall *et al.* noted that compared with traditional approaches, the key difference is that they keep track of the mean image $\bar{\mathbf{a}}_X \mathbf{1}$.

Similarly, a second eigenspace $\Theta(Y)$ is defined for the matrix Y, also of k rows but of p_Y columns. These eigenspaces are then added by concatenating the pair of collections, $Z = [X, Y]$.

$$\begin{aligned}
\Theta(X) &= (U_X, S_X, V_X, p_X, \bar{\mathbf{a}}_X) \tag{6.7} \\
\Theta(Y) &= (U_Y, S_Y, V_Y, p_Y, \bar{\mathbf{a}}_Y) \tag{6.8} \\
\Theta(Z) = \Theta(X) \oplus \Theta(Y) &= (U_Z, S_Z, V_Z, p_Z, \bar{\mathbf{a}}_Z) \tag{6.9}
\end{aligned}$$

The new eigenspace $\Theta(Z)$ is the addition of the column space and row space of X to the column space and row space of Y. Depending on the intersection of the eigenspaces the dimensionality is the same or less than the total dimensionality of $\Theta(X)$ and $\Theta(Y)$.

Using Hall *et al.*'s approach, the computational expense of adding two eigenspaces is $O(s^3)$, where s is the total number of non-zero singular values. For example, suppose X and Y are composed of 500 images but the SVD of both yields only 100 significant singular values. In this situation the computation time for the addition is $O(200^3)$, whereas computing the SVD for all 1000 images is $O(1000^3)$. Thus the technique allows for a very large training set constructed from smaller sets. A good summary of the technique is given in their final article [146].

6.8 Conclusion

This chapter has established that normal chest radiographs are clustered in a low dimensional sub-space of image space, with the cluster centred on the average image. Normal images tend to vary in a limited number of patterns. The magnitude of variance for each pattern is limited and for the sample used in the prototype system, it is indistinguishable from a normal (Gaussian) distribution.

The dimensionality of the cluster is difficult to estimate because there is no single method for determining which eigenimages are significant. However, for many applications this is unlikely to be a problem because higher order eigenimages represent patterns with near zero variance. Using the Guttman-Kaiser criterion the training set from the prototype system has 19 significant principal components. As the training set increases in size the number of significant components increases, but the rate of increase tends to slow as the set gets larger.

The training set of 77 normal images used for the prototype system is likely to have a sub-optimal number of images. It is sub-optimal because with this small set size, the effective rank increases rapidly when the number of training images increases. It would be preferable to have a set of at least 200–300 images, because with this larger set size it is hoped that the effective rank would be more stable. Despite the sub-optimal size of $p = 77$, the PCA process has identified 19 significant principal components (eigenimages)

Experiments suggest that for the high resolution images used in this thesis the dominant factor in computation speed is reading data from the disk, and this time increases linearly with increasing image size, i.e. increasing k. Thus much larger data images, such as 3-dimensional CT images, could be used with a linear increase in computation time. For chest radiographs it is possible to

perform PCA for training sets with several hundred images without changes to the method for the prototype system described in chapter 4.

Including pathology in the training set requires that the size of the training set be much larger because the training set must include enough examples of each pattern of variation for the eigenimages to be statistically significant. Given the extremely large number of pathologies identifiable on chest radiographs and the wide variety of appearances of each pathology, it is likely that an extremely large training set would be required.

In the last few years there has been significant work on methods of incrementally changing the training set, and more recently there has been work on methods of adding or subtracting two training sets. These techniques may allow eigenimage processing of the very large training sets that may be required for some radiological applications.

6.9 Summary

1. Chest radiographs are clustered within a multi-dimensional image space.

2. Chest radiographs tend to vary from the average in a limited number of ways. These are called principal components, $\mathbf{u}_i$.

3. The spread of images along each principal component is approximated by a Gaussian curve with variance λ_i.

4. The number of significant eigenimages is the effective rank of A. Using the Guttman-Kaiser criterion this is found to be $\epsilon = 14 - 19$ for the prototype system.

5. The effective rank varies with the number of images within the training set. While it does tend to stabilise, this is unlikely to occur until at least 200–300 chest radiographs, which is a larger training set than was feasible to acquire for the purpose of this thesis.

6. It is possible to develop systems where eigenspaces are added, allowing very large training sets to be constructed. This may be particularly relevant to

enable stabilisation of effective rank for other radiological images such as those from CT and MRI.

Chapter VII

Analysis of test images

This chapter studies the processing applied to the test images performed by the prototype system described in chapter 4. Initially this chapter presents a description of how each test image, $\mathbf{t}$, is decomposed into constituent images, as introduced in equation 4.5. That is, each test image is expressed as a linear combination of eigenimages plus a null space image. This chapter aims to provide both a visual interpretation and parameters for classification of both the column space image and the null space image, i.e. the portion of the test image described by the training set and the portion not described by the training set.

The work on visual interpretation of component images is motivated by earlier work on faces which noted that the first few eigenfaces tend to represent information such as lighting conditions and facial pose [147]. Other authors have shown that using knowledge of the visual interpretation of components allows adjustment of the relative importance of each component to improve classification accuracy [37]. The work on visual interpretation in this chapter has been presented at two image processing conferences [99, 148].

The test images used for this chapter are the test images from the prototype system, section 4.5. The set contains 30 known images: 15 normal and 15 abnormal.

7.1 Test images represented as linear combinations

The processing of a test image $\mathbf{t}$ involves projecting the image onto a new basis. Specifically the projection is onto the basis $\mathbf{u}_1 \ldots \mathbf{u}_k$, the orthonormal set of vectors that span the complete image space. For chest radiographs this complete image space is $k = 3 \times 10^6$ dimensions. Formally, the projection is expressed as

the linear combination

$$\mathbf{t} = \sum_{i=1}^{k} w_i \mathbf{u}_i. \tag{7.1}$$

The basis vectors $\mathbf{u}_1 \ldots \mathbf{u}_\epsilon$ are the retained principal components (eigenimages) of the training set, as initially found in the prototype system in section 4.4 and discussed in detail in chapter 6. The remaining components $\mathbf{u}_{\epsilon+1} \ldots \mathbf{u}_k$ are not specifically determined, but span the remainder of the complete image space not spanned by the principal components. With reference to the Singular Value Decomposition (SVD) of A, the vectors $\mathbf{u}_1 \ldots \mathbf{u}_\epsilon$ span the column space while $\mathbf{u}_{\epsilon+1} \ldots \mathbf{u}_k$ span the left null space. As noted previously, for most eigenimaging problems $\epsilon \leq p \ll k$.

In view of the distinction between the column space basis vectors and the left null space basis vectors, it is convenient to split equation 7.1 into two separate parts:

$$\begin{aligned}
\mathbf{t} &= \sum_{i=1}^{\epsilon} w_i \mathbf{u}_i + \sum_{i=\epsilon+1}^{k} w_i \mathbf{u}_i \tag{7.2} \\
&= \mathbf{c} + \mathbf{n} \tag{7.3}
\end{aligned}$$

The first part, $\mathbf{c}$, is the portion of the test image which lies in the column space of A. This image has the eigenimages, $\mathbf{u}_1 \ldots \mathbf{u}_\epsilon$, as its basis. The second part of the image, $\mathbf{n}$, lies in the left null space. Its basis vectors $\mathbf{u}_{1+\epsilon} \ldots \mathbf{u}_k$ are not calculated. As previously noted, for this thesis the part of a test image in the column space is referred to as the column space image and the portion in the left null space is referred to as the null space image. The null space image is always orthogonal to the column space image since a left null space is always orthogonal to the corresponding column space [26]. The separation of a test image into a column space image and a null space image is shown diagrammatically in figure 7.1. Similar to figure 6.1, image space is again represented by a 3-dimensional space $[\mathbf{a}_1,\mathbf{a}_2,\mathbf{a}_3]$, with the training set represented as a 2-dimensional elliptical disk with axes $\sqrt{\lambda_1}\mathbf{u}_1$ and $\sqrt{\lambda_2}\mathbf{u}_2$.

The separation of a test image into a column space image and a null space image allows for separate analysis of both parts. Section 7.3 discusses the information available from studying the column space image, while section 7.4
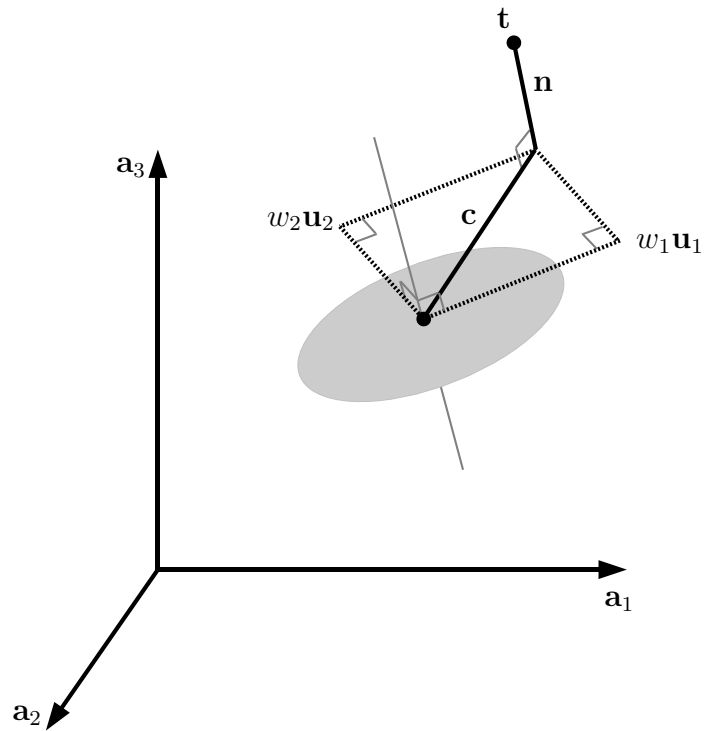
Figure 7.1: **Diagram of a test image separated into a column space image and null space image:** The diagram shows a test image $\mathbf{t}$ represented as a part that lies within the column space and a part that lies within the left null space, that is $\mathbf{t} = \mathbf{c} + \mathbf{n}$. Similar to diagram 6.1, the complete image space is 3-dimensional $[\mathbf{a}_1,\mathbf{a}_2,\mathbf{a}_3]$ while the training set sub-space is 2-dimensional (two principal components). The column space image $\mathbf{c} = w_1\mathbf{u}_1 + w_2\mathbf{u}_2$ lies in the plane of the training set, while the null space image $\mathbf{n}$ lies outside and orthogonal to training set sub-space. For clarity of the diagram, the column space image is shown to lie outside the control ellipse, but this will not necessarily be the case.

analyses the information available from the null space image. In chapter 8, a method for using information from both portions of the test image is presented.

7.2   Length of column space and null space images

At this point it is convenient to examine the length of the column space and null space images. From equation 7.2, any test image can be split into column and null space images. Since it is known that the column space image must be orthogonal to the null space image, Pythagoras' theorem can be used to relate the length of the images

$$\|\mathbf{t}\|^2 = \|\mathbf{c}\|^2 + \|\mathbf{n}\|^2 \tag{7.4}$$

By substituting equations 7.2 and 7.3 into equation 7.4, the lengths for the column space and null space images can established in terms of the weighting factors $w_1 \ldots w_\epsilon$. These lengths are

$$\|\mathbf{c}\| \;=\; \sqrt{\sum_{i=1}^{\epsilon} w_i^2} \tag{7.5}$$

$$\|\mathbf{n}\| \;=\; \sqrt{\|\mathbf{t}\|^2 - \sum_{i=1}^{\epsilon} w_i^2} \tag{7.6}$$

Using this simple method for calculating the length of the column space and null space images it is possible to determine if a test image can be well represented by a basis of eigenimages. The conjecture is that normal images should be well represented by the training set and therefore have a large column space component relative to the null space component. This method has been used by several authors for different applications of eigenimage processing [35,36,73,135].

For the 15 normal images, the ratio of the length of the null space image compared to the column space image is plotted in figure 7.2. This demonstrates that for many normal images the null space image is as large as the column space image. This again confirms that a training set of 77 normal images does not have enough images to accurately identify many common patterns of normal variance. This is in keeping with section 6.5 which postulated that the training set should contain at least 200–300 images.
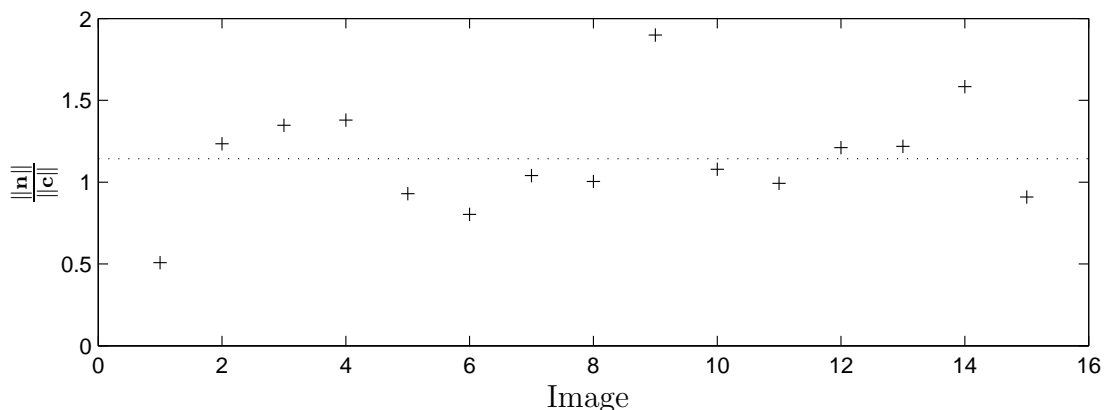
Figure 7.2: **Ratio of the length of null space and column space image for the 15 normal test images:** Crosses marks the ratio for each of the 15 normal test images. The dotted line is the average over the 15 images. The relatively large null space component for these normal test images suggests that the training set of 77 normal images contains too few images. If the training set better represented these 15 normal test images then the ratio would be smaller.

A plot of the ratio for abnormal images is not presented because it is likely that both component images will change when pathology is present, hence the two components are studied separately. Section 7.3 examines how the column space image changes when pathology is present, while section 7.4 examines how the null space image changes.

7.3   The column space image

This section discusses the analysis of the portion of a test image that lies within the column space $\mathcal{C}$. The column space is a sub-space of the complete image space. It has basis vectors given by the principal components of the training set.

$$\mathcal{C} = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_\epsilon\} \tag{7.7}$$

For image compression systems, dimensionality reduction problems, and Turk and Pentland's original face recognition system, this is the only portion of the test image considered [5, 6, 34, 40]. Initially a visual interpretation of the information

within each image is examined. Then an analysis of the weighting factors and how they might be used for classification are discussed.

7.3.1 Visual interpretation of column space images

As stated earlier, it has been shown by other authors that some components have more predictive value than others for representing certain features within a test image. The aim of this section is to visually investigate the eigenimages, with the hope of gaining a better understanding of what information they represent.

To visually study the column space images, they were reconstructed from their principal components, but using a filter that emphasises either low or high order components. When a signal is decomposed into its eigen-components and the relative importance of these components adjusted, the technique is referred to as eigen-domain enhancement. This is a common technique applied to MRI signals before image reconstruction, although it is less widely used elsewhere in image processing [149, 150].

To produce an image which enhances either the low or high order components the relative importance of each eigenimage is adjusted by introducing a filter parameter for each principal component. This is shown in equation 7.8 where an enhanced column space image, $\mathbf{c}_e$, is expressed as a linear combination of weighting factors w_i and eigenimages $\mathbf{u}_i$, but each component also has a filter parameter f_i corresponding to the ith eigenimage.

$$\mathbf{c}_e = \sum_{i=1}^{k} f_i \, w_i \, \mathbf{u}_i \qquad (7.8)$$

It is possible to define many types of filters acting in the eigen-domain. For this study two filters were chosen: "common pass" and "uncommon pass". These are analogous to high and low pass Fourier filters.

Common pass filters are defined as those in which low order eigenimages (e.g. $\mathbf{u}_1$ and $\mathbf{u}_2$) are weighted more heavily than high order eigenimages (e.g. $\mathbf{u}_{76}$ and $\mathbf{u}_{77}$). They are referred to as "common pass" because these eigenimages represent patterns of variance that are common within the training set of normal images.

Similarly, an uncommon pass filter is defined as one with high order eigenimages weighted more heavily than low order eigenimages. These high order
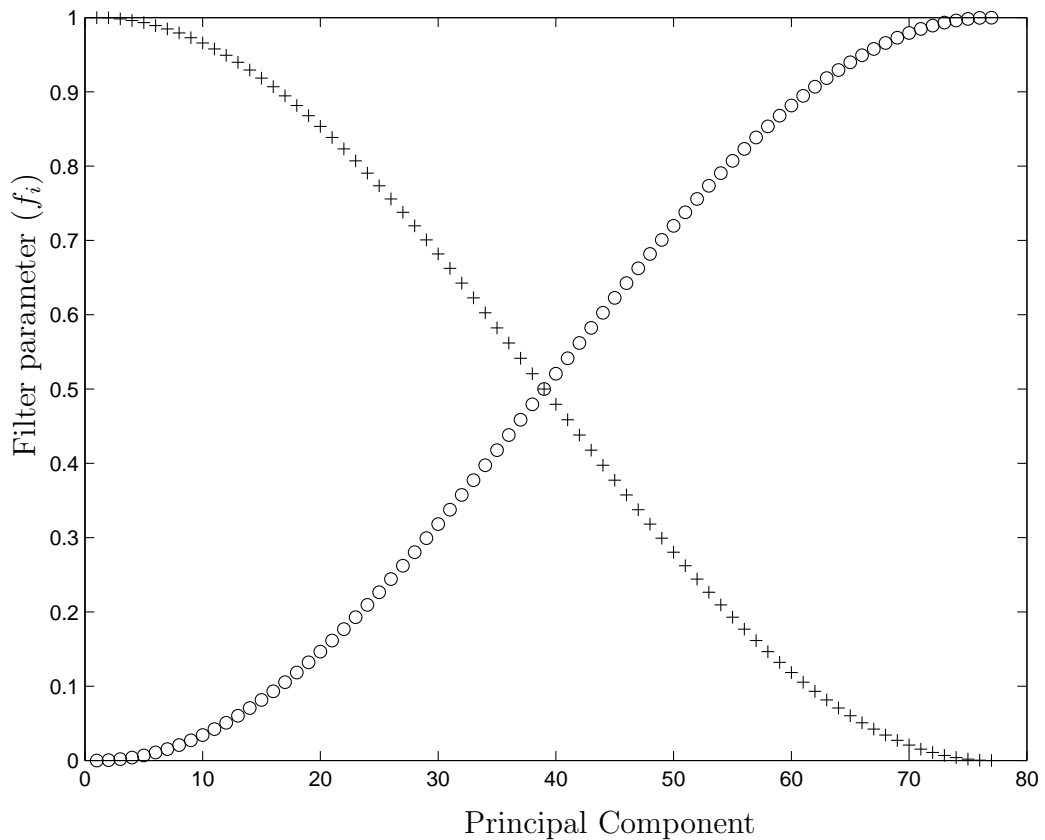
Figure 7.3: **Filter parameters for enhancing column space images:** The graph shows the chosen filter parameter f_i for common pass (+) and uncommon pass filters (o). Thus enhanced images visually show which components of a test image are represented by low order principal components and which parts are represented by high order principal components

eigenimages represent uncommon patterns of variance within the training set. For this study a cosine shaped filter was chosen, as graphed in figure 7.3. The choice of a cosine shaped filter was arbitrary and another shape could easily be substituted.

All 30 test images were enhanced in this manner. Figure 7.4 shows two examples from that test set: one normal and one showing right upper lobe pneumonia. Both images have undergone registration and intensity centring.

Examples of common pass filtered images are given in figure 7.5. As expected, common pass images tend to look like the original image because they

129

Normal image   Abnormal image

Figure 7.4: **Images to be enhanced in the eigen-domain:** The two images are examples of test images used to demonstrate enhancement in the eigen-domain. The abnormal image is an example of a right upper lobe pneumonia. Figures 7.5 and 7.6 show the same images after common pass and uncommon pass filtering.

are weighted towards components that are commonly found within the training set. Parts of the test image not represented by the training set of normal images are not visible. In the two example images, the right upper lobe pneumonia is a pattern of variance not represented within the training set, and hence the pattern for this pneumonia is not visible.

Two examples of uncommon pass filtered images are shown in figure 7.6. These images tend to look more noisy as they are constructed from patterns of variance that are uncommon within the training set. Again the pathological finding of right upper lobe pneumonia is not clearly visible in the filtered image because that pattern of variance is not found within the training set.

In summary, while the common and uncommon pass images contain different information, neither is good at showing pathology. This is thought to be because pathological features are not contained within the training set and therefore cannot be represented by the eigenimage basis.
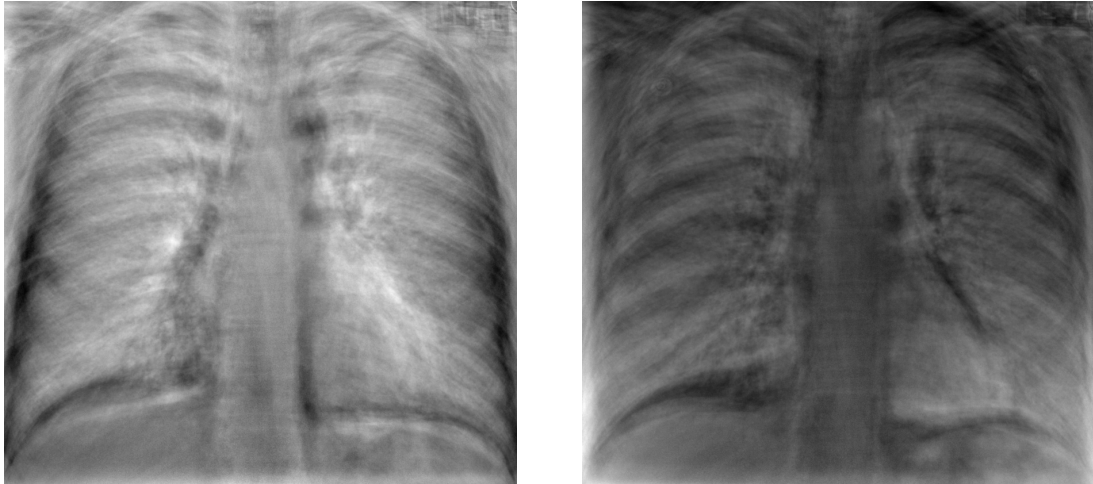
Normal image   Abnormal image

Figure 7.5: **Images after common pass filtering:** The two images are the radiographs from figure 7.4 after processing by the common pass filter. Note how the right upper lobe pneumonia, which should be present in the abnormal image, is barely visible.

7.3.2   Classification using the column space image

The aim of this section is to improve on the Euclidean length metric that was used for classification in the prototype system. The hope is that a better metric may be obtained by comparing the location of a test image to the multivariate confidence interval, the control ellipse, introduced in chapter 2. For $\epsilon > 2$ the hyper-dimensional analogue is a control ellipsoid.

Although section 7.3.1 showed the column space image does represent pathology well, it is known that a difference in the column space between normal and abnormal images does exist. In the prototype system, described in chapter 4, this was demonstrated by the finding that classification was successfully performed using the Euclidean length of weighting factors alone (figures 4.6 and 4.7).

Equation 7.2 and 7.3 express the column space image in terms of the weighting factors and the eigenimages

$$\mathbf{c} = \sum_{i=1}^{\epsilon} w_i \mathbf{u}_i. \tag{7.9}$$

131

| Normal image | Abnormal image |

Figure 7.6: **Images after uncommon pass filtering:** The two images are the radiographs from figure 7.4 after processing by the uncommon pass filter. Again, note how the right upper lobe pneumonia, which should be present in the abnormal image, is barely visible.
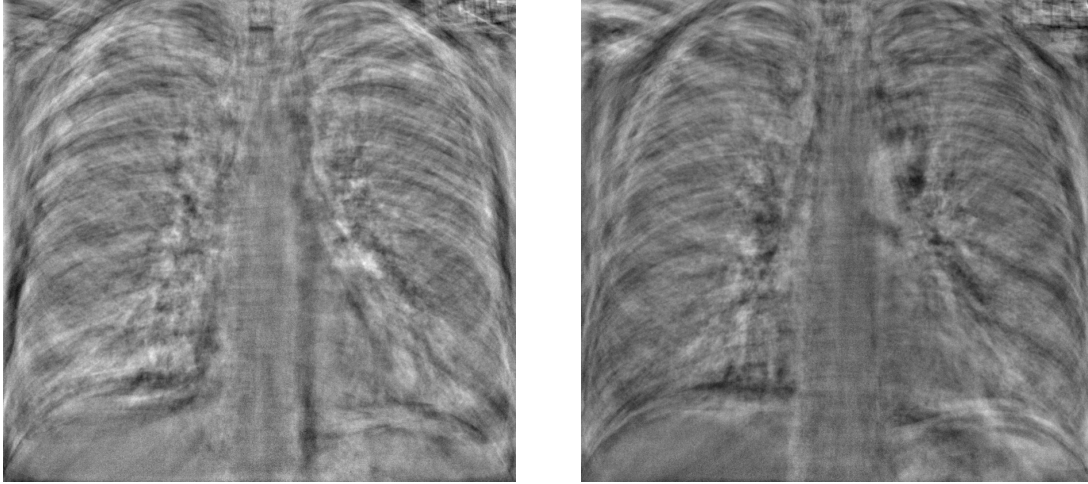
As discussed, the column space image lies within $\mathcal{C}$, an ϵ-dimensional space defined by the basis vectors of the eigenimages. In addition, section 6.4 demonstrates that the variance along each principal component is of limited magnitude, given by the associated characteristic root, λ_i. In both figures 2.3 and 6.1, the amount of variance associated with each characteristic root is represented by the control ellipse.

To construct the multi-dimensional confidence interval (the control ellipsoid), first note that any point $\mathbf{g}$ in the column space can be expressed in terms of the eigenimage basis

$$\mathbf{g} = \sum_{i=1}^{\epsilon} g_i \mathbf{u}_i. \tag{7.10}$$

To ensure that $\mathbf{g}$ lies on the surface of an ellipsoid, g_i is constrained by

$$\sum_{i=1}^{\epsilon} \frac{g_i^2}{\left(\sqrt{T_{\epsilon,p,\alpha}^2 \lambda_i}\right)^2} = 1 \tag{7.11}$$
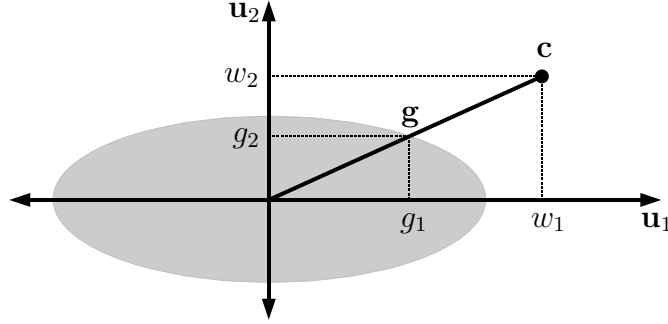
Figure 7.7: **Diagram of a control ellipse:** The ellipse is the disc in figure 7.1. In ϵ dimensions, where $\epsilon > 2$, the ellipse's analogue is an ellipsoid. In the diagram the axes $\mathbf{u}_1$ and $\mathbf{u}_2$ are the two principal components. $\mathbf{c}$ is the column space image, which can be represented as $\mathbf{c} = w_1\mathbf{u}_1 + w_2\mathbf{u}_2$. The point where the vector $\mathbf{c}$ intersects the ellipse is $\mathbf{g} = g_1\mathbf{u}_1 + g_2\mathbf{u}_2$.

where $\sqrt{T^2_{\epsilon,p,\alpha}\lambda_i}$ is the radius of the ellipsoid along each eigenimage and $T^2_{\epsilon,p,\alpha}$ is Hotelling's T^2 distribution, introduced in section 2.6.

The parameters of $T^2_{\epsilon,p,\alpha}$ may be chosen as: $\alpha = 0.95$ (95% control ellipsoid), $p = 77$ (number of observations), $\epsilon = 57$ (independent patterns of variance, i.e. retained components). For the data presented here, the mean and co-variance matrices have been found by sampling and may contain outliers. Hence the β-distribution is used to calculate T^2 (see equation 2.57). Choosing to retain all $\epsilon = 77$ components is not possible because the β-distribution from which the T^2 distribution is derived is not defined for $\epsilon = p$.

Figure 7.7 shows a two-dimensional representation of the sub-space formed by the principal components of the training set. The column space image $\mathbf{c}$ is shown in relation to the control ellipse. The point $\mathbf{g}$ is where the vector of the column space image intersects the control ellipse. Formally, the vector from the origin to $\mathbf{g}$ and the vector from the origin to $\mathbf{c}$ are co-linear. Specifically this is

$$r_\alpha\,\mathbf{g} = \mathbf{c}, \tag{7.12}$$

where r_α is a scalar quantity relating the column space image to control ellipse of width α. If $r_\alpha \leq 1$ then the column space image is either on or inside the control ellipse. However, if $r_\alpha > 1$ then the column space image is outside the control

ellipse. Thus, the scalar r_α provides a natural measure for determining whether the weighting factors of a test image fall within the expected normal range, as determined by the characteristic roots.

Substituting equations 7.9 and 7.10 into equation 7.12 gives

$$\sum_{i=1}^{\epsilon} r_\alpha \, g_i \, \mathbf{u}_i = \sum_{i=1}^{\epsilon} w_i \, \mathbf{u}_i. \tag{7.13}$$

Since the $\mathbf{u}_i$ are orthonormal, the corresponding terms on both sides of equation 7.13 can be equated yielding

$$r_\alpha \, g_i = w_i \qquad \text{for } i = 1 \dots \epsilon, \tag{7.14}$$

which is consistent with the geometry in figure 7.7. Substituting $g_i = w_i/r_\alpha$ into equation 7.11 gives

$$r_\alpha = \frac{1}{T_{\epsilon,p,\alpha}} \sqrt{\sum_{i=1}^{\epsilon} \frac{w_i^2}{\lambda_i}}. \tag{7.15}$$

This new control ellipsoid metric can be compared to the Euclidean distance metric used for the prototype system (equation 4.7). Both metrics have a very similar form, except that the control ellipsoid metric scales the weighting factors w_i by the variance along each component λ_i.

The control ellipsoid metric should also be compared to the Mahalanobis distance introduced in section 2.5. It can be seen that the control ellipsoid metric is the Mahalanobis distance of the column space image from the average image, but scaled by T^2. That is, r_α is the Mahalanobis distance from the column space image to the average, relative to the expected distribution

$$r_\alpha = \frac{D_M(\mathbf{c})}{T_{\epsilon,p,\alpha}^2}. \tag{7.16}$$

Using the characteristic roots from the prototype system, the control ellipsoid metric $r_{0.95}$ was calculated for both the 15 normal and 15 abnormal test images. Cumulative probability distributions for both the normal and abnormal images are plotted in figure 7.8. Studying the distributions it can be seen that both the normal and abnormal images have column space images which tend to lie
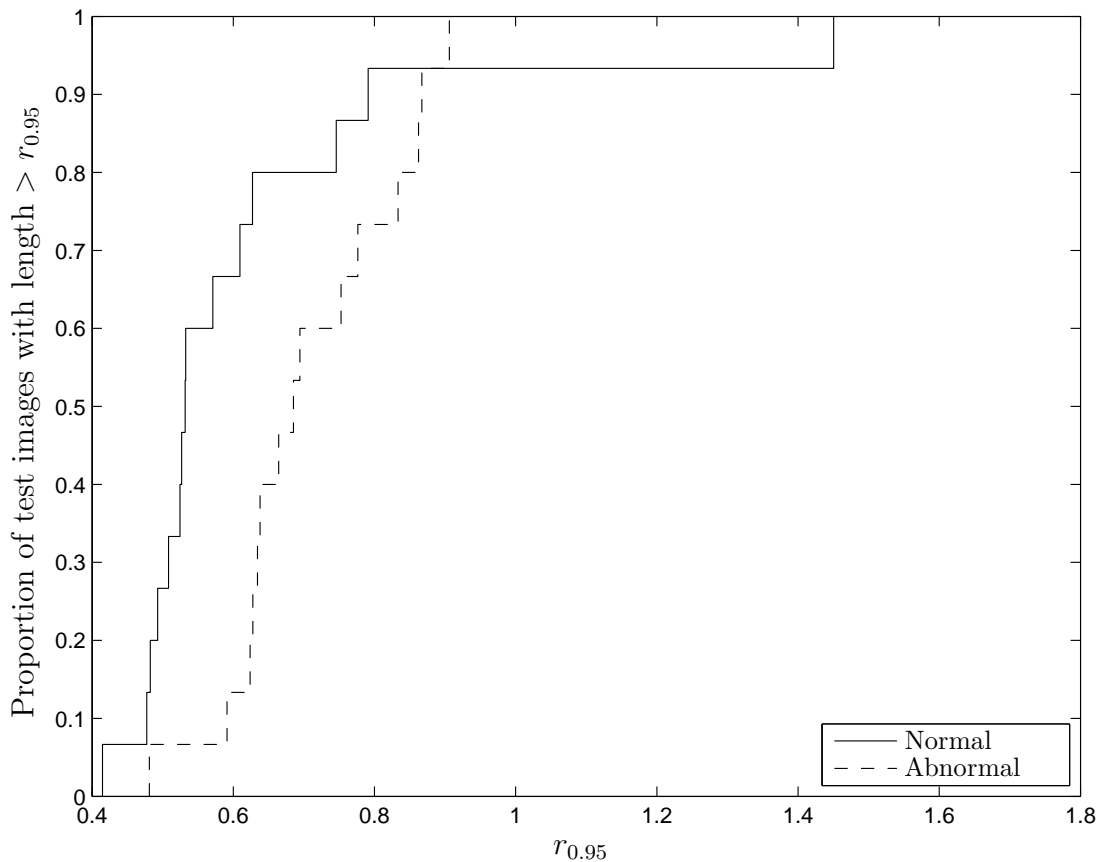
Figure 7.8: **Control ellipsoid metric:** Plot of the cumulative probability distribution for $r_{0.95}$. The metric is based on the relative distance of each column space image to the 95% control ellipsoid.

within the 95% control ellipsoid, although normal images tend to be closer to the average image. That is, $r_{0.95}$ tends to be smaller for normal images than abnormal images.

If the cumulative probability distributions of the control ellipsoid metric and the Euclidean length metric (cf. figure 4.6) are compared, it can be seen that for the control ellipsoid metric the distributions are more separated than for the Euclidean length metric. This suggests that the control ellipsoid metric is likely to be better for classification than the Euclidean length metric. One method for comparing the predictive value for both metrics is to compare the ROC curves. Figure 7.9 plots ROC curves using both metrics. It confirms that the ellipsoid

metric is a better classification metric for the column space image.

The area under a ROC curve (AUC) measures a test's discrimination [85, 86]. A test's discriminatory power is the probability that, if a normal and an abnormal test image are chosen, the image with the most abnormal metric will be correctly classified as an abnormal image. Using the example in figure 7.9, if both a normal and abnormal test image have their Euclidean length metrics calculated, there is a 0.67 probability that the abnormal test image would have the largest Euclidean metric. While discriminatory power is important in determining the clinical utility of a test, the shape of the ROC curve is also often very important because the clinical requirements could be for high specificity or high sensitivity, rather than simply high overall discriminatory power.

For the tests in figure 7.9, the AUC for the Euclidean distance metric was 0.67, while the AUC for the control ellipsoid metric was 0.79, suggesting that the control ellipsoid metric produces a test with better discriminatory power.

In summary, while the column space image does not contain all the information regarding pathology, it is possible to use the weighting factors for classification. The location relative to a control ellipsoid produces a metric which is of more predictive value than the Euclidean length metric used for the prototype system. This result is consistent with the finding by several previous authors that Mahalanobis distance can produce better classification than the Euclidean distance for face recognition systems [55, 56, 151].

7.4   The null space image

This section studies the null space image $\mathbf{n}$. This is the part of a test image that lies within the left null space $\mathscr{N}$ and is therefore not described by the retained set of ϵ eigenimages. The left null space is given by,

$$\mathscr{N} = \{\mathbf{u}_{\epsilon+1}, \, \mathbf{u}_{\epsilon+2}, \, \ldots, \, \mathbf{u}_k\}, \tag{7.17}$$

although for the processing performed in this thesis these basis vectors are not explicitly calculated.

Study of the null space is described in Jackson's text on Principal Components Analysis (PCA) where it is referred to as "residual analysis" [25]. With respect
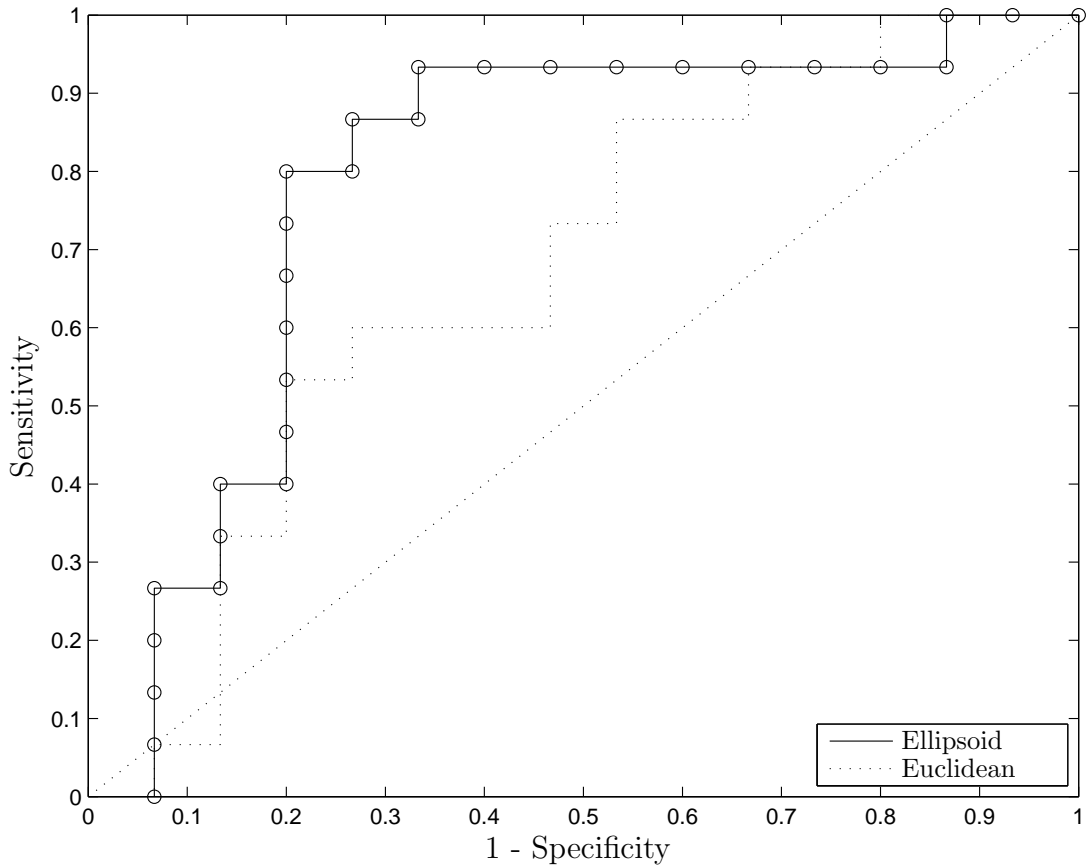
Figure 7.9: **Receiver operating characteristic curve for control ellipsoid metric:** The ROC curve for classification using the control ellipsoid metric and Euclidean length are plotted together for comparison. The control ellipsoid metric produces a curve that is significantly further towards the upper left region of the graph, thus it provides better sensitivity and specificity than the Euclidean length metric. Both ROC curves here are produced using $\epsilon = 57$, whereas for the prototype system all $\epsilon = 77$ principal components were retained. The AUC values are 0.67 for the Euclidean metric and 0.79 for the ellipsoid metric.

to eigenimage processing, Vaswani and Chellapa published an article in 2006 on null space analysis for image and video classification systems [152].

Null space analysis is infrequently used within eigenimage processing because if a test image is close to the training set, then most of the test image is represented by the eigenimages, i.e. the null space image is small. For the face recognition system described by Turk and Pentland, all new faces should be well described by the training set, hence the null space is unlikely to be useful. However, for the prototype system described in chapter 4, the training set only includes normal radiographs. Hence for a test image containing pathology or a rare normal variation the null space image may be large. This makes null space analysis worthwhile for processing of radiological images.

To calculate the null space image, each test image has its column space image removed, thus

$$\mathbf{n} = \mathbf{t} - \sum_{i=1}^{\epsilon} w_i \mathbf{u}_i. \tag{7.18}$$

The calculation is performed in this way because the basis vectors for the null space are not found during the PCA process. For the data presented in this section, the eigenimages are those from the prototype system. The effective rank was chosen to be $\epsilon = 57$ so that the eigenimages span 95% of the total variance.

7.4.1 Visual interpretation of null space images

Once again the technique of visual interpretation has been used to assess null space images, in order to gain insight into features that may be present. Null space images are displayed using absolute pixel values. That is, the brightness of a pixel is proportional to the absolute (unsigned) value of the pixel. The images are displayed as absolute images because this illustrates how far the test image is away from the sub-space defined by the normal training set, for each dimension in the image space (i.e. each pixel). Two examples of null space images are given in figure 7.10.

It is also possible to show null space images on a grey scale from black (minimum) to white (maximum). For those pixels where the null space component is zero, the pixel is displayed as grey. Pixels where the null space component is negative are displayed as darker shades, while pixels where the remainder is
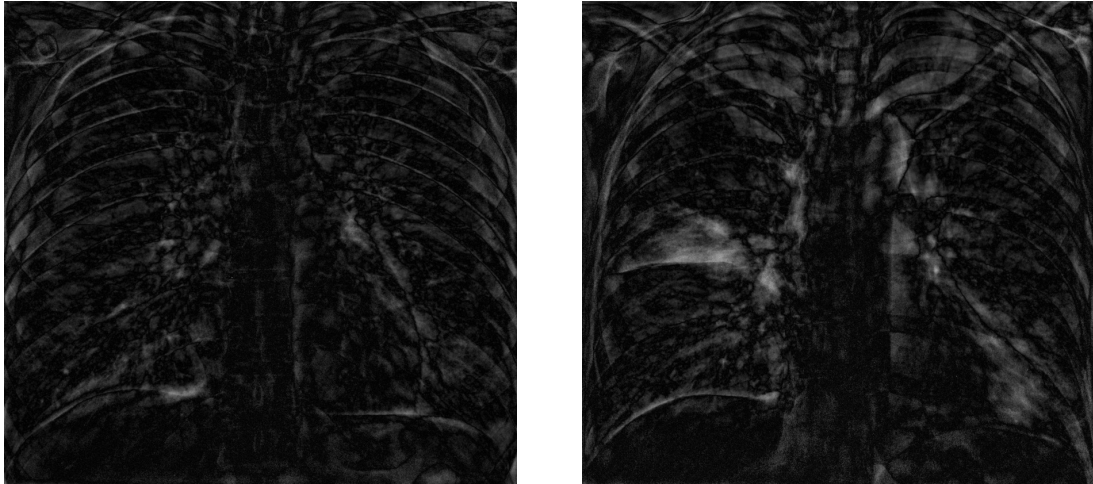
Normal image   Abnormal image

Figure 7.10: **Null space images:** The images are displayed with intensity equal to the absolute value. The two images correspond to those from 7.4. The normal test image produces an null space image that is generally of low intensity. The abnormal test image produces a null space image where there is a clearly noticeable area of higher intensity corresponding to the right upper lobe pneumonia. Note that both images are displayed using the same intensity scale.

positive are displayed as lighter shades. Example of these grey scale images are in figure 7.11, although the author tends to find the absolute images are more visually striking.

The null space images in figures 7.10 and 7.11 demonstrate that the null space image derived from a normal test image is of low intensity. This is expected because normal test images should be well represented by the basis of eigenimages derived from the normal training set. For the abnormal test image, the null space image has a region of high intensity where the right upper lobe pneumonia is located. This is expected because the right upper lobe pneumonia is unlikely to be a pattern of variance found in the normal training set, hence there are no eigenimages present which are able to represent the pattern.
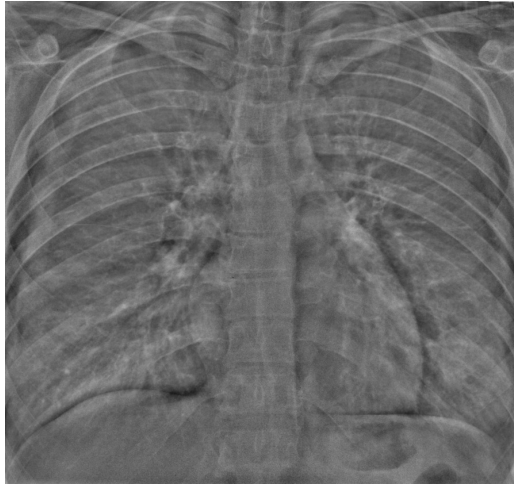
| Normal image | Abnormal image |

Figure 7.11: **Null space images displayed using grey scale:** The two null space images from figure 7.10 are displayed using a grey scale, rather than the absolute value.

7.4.2 Classification based on the null space

Having established that the null space image is likely to contain at least some patterns representing pathology, it should be possible to perform classification based on the null space image alone. However, there are few precedents within image processing to determine which metric should be used for this classification process. Using the geometry depicted in figure 7.1 a reasonable metric to use for null space image classification is the length of the null space image.

Within the framework of PCA, the term residual analysis has been applied to the study of null space information. Residual analysis is performed using the Q-statistic or Q-metric which was introduced in 1957 to complement the use of control ellipses [25, 153] by analysing the variance not explained by the principal components.

It transpires that the square of the Euclidean length of a null space image is

identical to the Q-metric.

$$Q(\mathbf{n}) \;=\; \sum_{i=1}^{k} n_i^2 \qquad\qquad (7.19)$$

$$\;=\; \mathbf{n}^T \mathbf{n} \qquad\qquad (7.20)$$

$$\;=\; D_E(\mathbf{n})^2 \qquad\qquad (7.21)$$

where n_i is the intensity of the ith pixel. Thus there are two reasons to consider equation 7.19 a good metric for comparing null space images.

Within image processing this metric is sometimes known as "image energy". The visual correlate of the Q-metric is image brightness. That is, for the absolute null space images shown in figure 7.10, images with a large Q have more bright pixels.

In PCA's residual analysis, estimates of the distribution of Q are based on all characteristic roots being calculated, but without all principal components being retained. The characteristic roots from the non-retained components are used to determine the maximum expected Q. That is, $\lambda_{\epsilon+1} \ldots \lambda_k$ are used to determine a maximum expected value.

Unfortunately, for eigenimage processing, while roots $\lambda_{\epsilon+1} \ldots \lambda_p$ are found, roots $\lambda_{p+1} \ldots \lambda_k$ are not. This is because the sample co-variance matrix is made from fewer observations than there are variables ($p \ll k$). Thus, the methods for estimating a maximum expected value for Q cannot be used.

Despite not being able to estimate a maximum expected value for Q, there is still a good reason to choose it as a metric for performing image classification. Fortunately, it is possible to measure Q for both sets of normal and abnormal images, thus estimating Q's distribution. Figure 7.12 shows an analysis of null space image brightness Q for the 15 normal and 15 abnormal test images. The plot confirms that abnormal images tend to be brighter than normal images, i.e. abnormal images tend to have larger Q value.

Using the Q-metric it is possible to classify images as normal or abnormal. This is performed in a similar manner to the classification based on column space metrics. A ROC curve for this classification is shown in figure 7.13. The ROC curve lies above and to the left of the diagonal line representing no predictive value, and it has an AUC of 0.74. These confirm that classification based on the
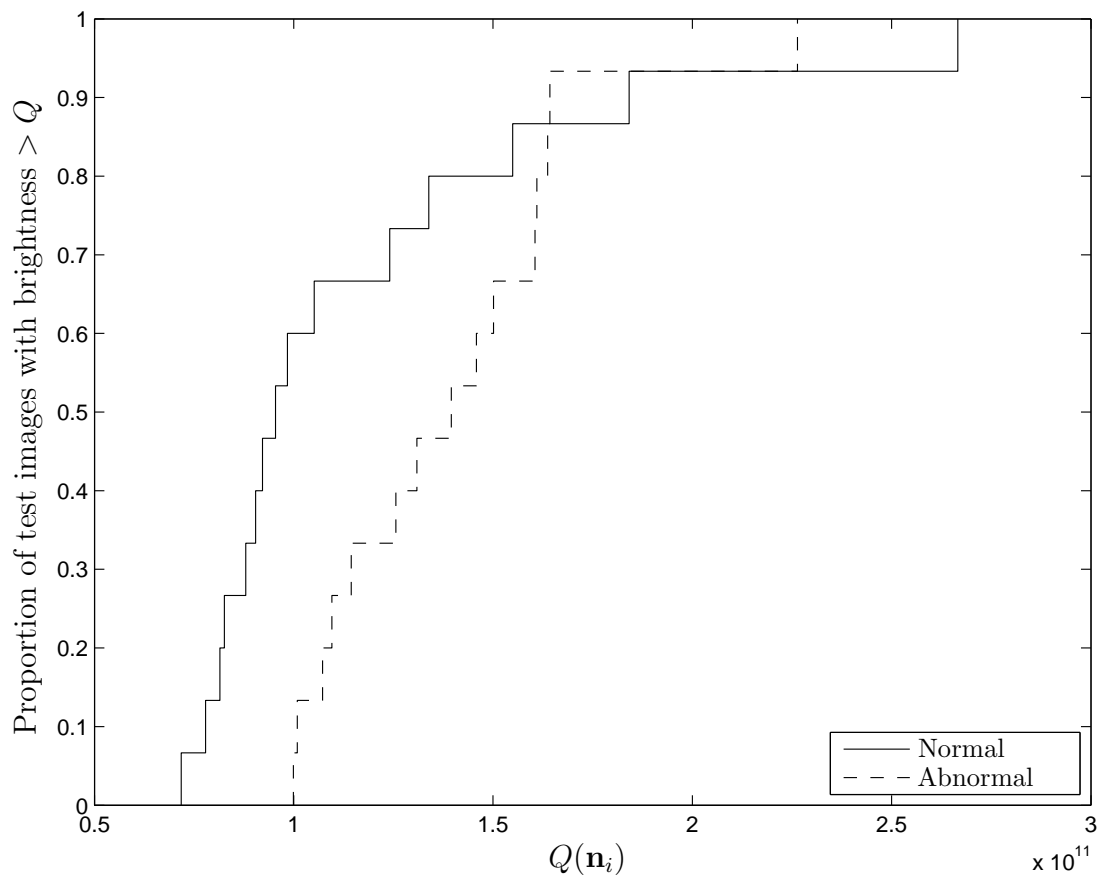
Figure 7.12: **Null space image brightness:** Plot of the cumulative probability distributions of Q for normal and abnormal test images. The plot demonstrates that abnormal images tend to be brighter than normal images.

Q-metric does have predictive value.

7.5   Test image analysis with a mixed training set

In section 6.3.4 an experiment is described which aimed to demonstrate how adding pathological images to the training set alters the eigenimages and characteristic roots. That section concludes that such a training set would need to be very large and should include a wide range of pathologies. Section 6.7 discussed how such a training set might be constructed by adding smaller sets together.
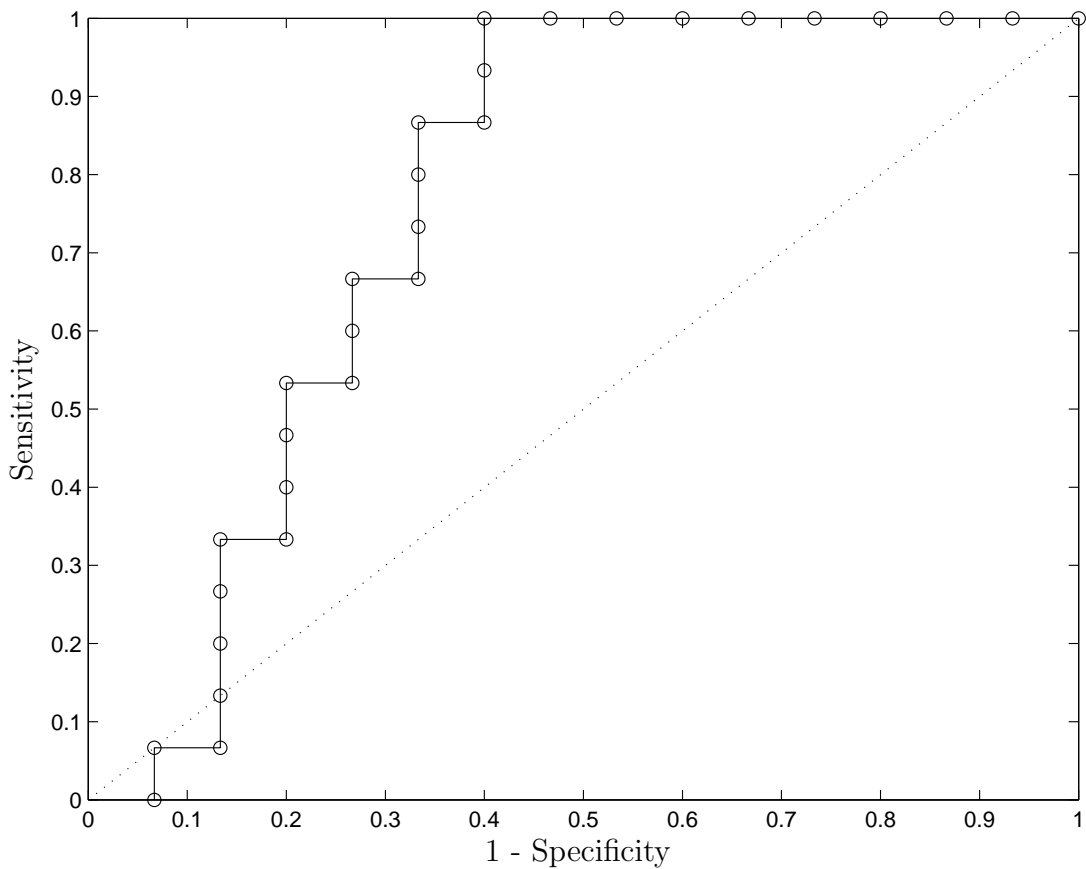
Figure 7.13: **Receiver operating characteristic curve for classification using the null space image:** Using the Q-metric classification is performed to assign test images as either normal or abnormal. The AUC for this test is 0.74.

Unfortunately, gathering enough images and developing appropriate routines to do this is beyond the scope of this thesis. Despite being limited to training sets containing only normal images, it is still worth considering how the analysis of test images might be affected by a training set that contained a mix of normal and pathological images.

If the training set were to include a wide range of pathologies, the column space of A would contain patterns that represent pathology. Thus, in addition to eigenimages that represent normal findings, there would be eigenimages which represent patterns of pathology. The centre of the training set cluster would

not be "the average normal image", but rather it would simply be "the average image".

The classification scheme would need to identify the regions in eigenspace that correlate to certain pathologies. It is not known if pathologies would cluster in eigenspace, but there is reason to predict that similar pathologies could be very spread out. For example, left lung pneumonia is unlikely to be close in eigenspace to right lung pneumonia. The problem gets worse for diseases such as lung cancer, which can occur anywhere within the lung. As mentioned in the introduction to eigenimage processing, section 1.5, one method of coping with this is to define separate sub-spaces for each class (i.e. pathology) to be identified and then determine which sub-space a test image is closest to [45, 57, 58].

Understanding the effect on the null space image is simpler than understanding the effect on the column space image. With a mixed training set the eigenimages would be able to better represent pathology, hence there would be less information contained in the null space image. Studying the size of the null space could be used to determine if the training set contained enough forms of pathology.

In summary, if pathological images were to be mixed into the training set, the classification systems developed in this chapter would need to be altered. A good starting point for this would be to measure the distance between the cluster of normal images and a cluster associated with each class of pathology. De Maesschalck describes a method for this using Mahalanobis distance [69].

7.6 Conclusion

Both the column space and null space images contain information that has predictive value for classification. Therefore, an ideal system for classification or enhancement of chest radiographs would incorporate information provided by both the column space image and null space image.

It is known that the column space does contain information useful for classification, because weighting factors tend to be larger for abnormal images than normal images. That is, abnormal column space images tend to be further from the centre of the cluster of normal images found using PCA. A good metric for measuring this distance from the centre of the normal cluster is Mahalanobis

distance, which is a good metric because it takes into account the variance associated with each principal component.

Visual demonstrations show that pathology is more likely to appear in the null space image than the column space image. This is thought to be because abnormal patterns cannot be well represented by basis eigenimages formed from a normal training set. Viewing null space images is helpful because pathology is often demonstrated. Classification based on the null space image can be performed using the Q-metric which visually correlates to image brightness.

Previous experiments to determine effective rank were inconclusive (section 6.3). They suggested ϵ could range from 14 to 57 and thus there remains uncertainty as to the effective rank that would provide maximum predictive value. While the choice of effective rank could be optimised separately for the column space and null space classification system presented in this chapter, it will be optimised following implementation of a complete system that combines the column space and null space information.

Chapter 6 concluded that a training set of 77 normal images was likely to be too small. When studying the relative lengths of the null space and column space images, it was seen that for normal test images the null space image is large proportion of the image. Again, this confirms that the training set is too small to accurately represent the large number of ways in which normal chest radiographs may vary.

7.7   Summary

1. The column space images contain patterns of variance found within the training set.

2. Weighting factors associated with each eigenimage tend to be larger for abnormal test images compared to those for normal test images.

3. To improve classification, the importance of each weighting factor should be adjusted relative to the control ellipsoid of the normal training set, by using Mahalanobis distance.

4. The null space image shows patterns of variance not found within the training set.

5. The length (or brightness) of the null space image, the Q-metric, can be used to predict whether a test image contains pathology.

6. Comparing the lengths of the null space and column space image for normal test images suggests that the training set is too small. This is in keeping with findings from chapter 6.

7. The column space and null space images both contain diagnostic information. An ideal classification or enhancement routine should incorporate information from both images.

8. Following design of a system that incorporates both column space and null space information, the effective rank needs to be optimised to maximise discrimination.

Chapter VIII

The remainder image

This chapter develops and investigates a new derived image, referred to as the "remainder image", $\mathbf{r}$. The remainder image is formed from a test image by subtracting the component of the test image which falls within a control ellipsoid that circumscribes the cluster of normal training images. This is a novel technique developed for this thesis.

The decision to derive the remainder image is based on the work on null space analysis presented in section 7.4. That work concluded that when using a training set of normal images, a good method for classification is to study the "part of an image not represented by the training set". In that section, the part not represented by the training set was interpreted to be solely the null space image. However, considering the null space image alone is too simplistic because for a pathological test image the column space image is likely to be further from the centre of the control ellipsoid than is the case for normal images. Thus there are two ways in which a test image can be "not represented by the training set": in its column space component, and in its left null space component. The remainder image attempts to take into account information from both the column space and null space images.

This chapter introduces the remainder image both diagrammatically and mathematically. Following this, a classification routine is presented, based on the Q-metric introduced for null space classification schemes.

For any control ellipsoid used to circumscribe a cluster of normal training images there are two parameters, α and ϵ, which adjust the span of the ellipsoid. These parameters adjust the dimensionality (effective rank) and width (proportion of the cluster spread to be spanned) of the control ellipsoid. Section 8.3.2 optimises these two parameters to provide a maximum predictive value for the classification routine.

Finally the chapter visually presents selected test images and their associated remainder images. These images demonstrate that the remainder image can be used as a form of image enhancement as well as in classification routines.

8.1 Definition of the remainder image

The remainder image is defined as the component of any test image that lies outside the control ellipsoid. Two separate situations occur:

The column space image is inside the control ellipsoid ($r_\alpha \leq 1$): In this situation the only part of the test image not represented by the cluster of normal training images is the portion orthogonal to the plane of the control ellipsoid (i.e. the null space image). This is represented in the left diagram of figure 8.1.

The column space image is outside the control ellipsoid ($r_\alpha > 1$): In this situation the test image is outside the cluster of normal images in two separate respects. Firstly it has a part that lies orthogonal to the plane of the control ellipsoid (i.e. the null space image). Secondly its location on the plane of the control ellipsoid is further from the average image than expected. In this situation, the remainder image is the vector from the closest point on edge of the control ellipsoid to the test image. This is represented in the right diagram of figure 8.1.

This definition ensures that the remainder image contains all the information about the component of a test image outside the control ellipsoid. Specifically it includes information from both the column space component and the null space component.

The two situations are represented mathematically by

$$\mathbf{r} = \begin{cases} \mathbf{t} - \mathbf{c} & \text{for } r_\alpha \leq 1, \\ \mathbf{t} - \mathbf{g} & \text{for } r_\alpha > 1. \end{cases} \tag{8.1}$$

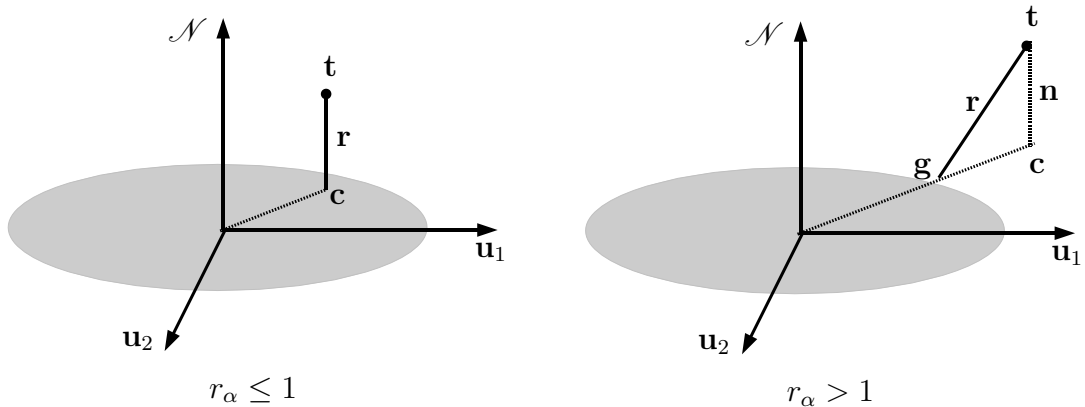Both r_α and $\mathbf{g}$ were introduced in section 7.3.2.

Figure 8.1: **Diagrams of remainder images:** These diagrams illustrate the two possible cases for the remainder image **r**. In the left diagram the column space image lies inside the control ellipsoid ($r_\alpha \leq 1$). In this situation, the remainder image is the part of the test image that is orthogonal to the plane of the control ellipsoid. In the right diagram the column space image is outside the control ellipsoid ($r_\alpha > 1$). In this situation the remainder image is the vector between the edge of the control ellipsoid at location **g** and the test image. As in previous diagrams, the control ellipsoid is here represented as a two dimensional ellipse on the plane defined by the basis vectors for the column space $\mathcal{C}$; $\mathbf{u}_1$ and $\mathbf{u}_2$. The left null space $\mathcal{N}$ is represented as a single dimension orthogonal to the plane of the column space. For eigenimage processing the dimensionality of the null space is $k - \epsilon$, with basis vectors $\mathbf{u}_{\epsilon+1} \ldots \mathbf{u}_k$.

8.2 Calculating the remainder image

For simplicity, the method for calculating the remainder image will initially be separated into the two situations described in section 8.1. First presented will be the case where the column space image is within the control ellipsoid. Next will be the case of the column space image being outside the control ellipsoid. Finally the two expressions for the remainder image are combined to give a single simpler representation.

The control ellipsoid metric r_α is used to decide whether the column space image **c** lies within the control ellipsoid. The position of the column space image

relative to the edge of the control ellipsoid is given by r_α. It was developed in section 7.3.2, and expressed in equation 7.15.

If $\mathbf{c}$ lies within or on the control ellipsoid, the remainder image is the test image minus the entire column space image.

$$\mathbf{r} \;=\; \mathbf{t} - \mathbf{c} \qquad\qquad \text{for } \; r_\alpha \leq 1 \qquad\qquad (8.2)$$

$$\;=\; \mathbf{t} - \sum_{i=1}^{\epsilon} w_i \mathbf{u}_i. \qquad\qquad\qquad (8.3)$$

This is identical to the null space image from section 7.4.

If the column space image lies outside the control ellipsoid, the remainder image is the test image minus the nearest approximation to the test image which lies inside the control ellipsoid. If $\mathbf{g}$ is the point at which the vector from the origin to $\mathbf{c}$ intersects with the control ellipsoid (section 7.3.2), the remainder image is given by

$$\mathbf{r} \;=\; \mathbf{t} - \mathbf{g} \qquad\qquad \text{for } \; r_\alpha > 1 \qquad\qquad (8.4)$$

$$\;=\; \mathbf{t} - \sum_{i=1}^{\epsilon} g_i \mathbf{u}_i \qquad\qquad\qquad (8.5)$$

where g_i is found from $g_i = w_i / r_\alpha$ (as per equation 7.14). This leads to

$$\mathbf{r} = \mathbf{t} - \frac{1}{r_\alpha} \sum_{i=1}^{\epsilon} w_i \mathbf{u}_i. \qquad\qquad\qquad (8.6)$$

Thus, the remainder image can be calculated for the two situations in figure 8.1

$$\mathbf{r} = \begin{cases} \mathbf{t} - \sum_{i=1}^{\epsilon} w_i \mathbf{u}_i & \text{for } r_\alpha \leq 1 \\ \mathbf{t} - \frac{1}{r_\alpha} \sum_{i=1}^{\epsilon} w_i \mathbf{u}_i & \text{for } r_\alpha > 1 \end{cases} . \qquad\qquad (8.7)$$

This can be simplified into a single expression

$$\mathbf{r} = \mathbf{t} - \frac{1}{r'_\alpha} \sum_{i=1}^{\epsilon} w_i \mathbf{u}_i \begin{cases} r'_\alpha = 1 & \text{for } r_\alpha \leq 1 \\ r'_\alpha = r_\alpha & \text{for } r_\alpha > 1. \end{cases} \qquad (8.8)$$

8.3 Classification using the length of the remainder image

If the remainder image contains information regarding how "far from normal" a test image is, then it should be possible to use the remainder image for classification. It is possible that the remainder image is better for classification than either the null space or column space images alone, as it contains information from both.

8.3.1 The Q-metric for remainder images

In order to perform classification, a metric must be developed to allow comparison of two remainder images. The most obvious metric to use is the Q-metric introduced in section 7.4.2. This is the square of the length of $\mathbf{r}$:

$$Q_r = \mathbf{r}^T \mathbf{r} = \|\mathbf{r}\|^2 \tag{8.9}$$

The larger Q_r is, the "further away" a test image is from the cluster of normal images.

Interestingly, Q_r can be calculated in two ways: either by finding $\mathbf{r}$, or by using several derived variables. The second method is useful because in some situations it may be computationally faster than calculating $\mathbf{r}$. To develop this faster technique, consider an expansion of Q_r,

$$Q_r = \|\mathbf{r}\|^2 \tag{8.10}$$

$$= \left\| \mathbf{t} - \frac{1}{r'_\alpha} \sum_{i=1}^{\epsilon} w_i \mathbf{u}_i \right\|^2 \tag{8.11}$$

$$= \|\mathbf{t}\|^2 + 2 \left\langle \mathbf{t}, \frac{-1}{r'_\alpha} \sum_{i=1}^{\epsilon} w_i \mathbf{u}_i \right\rangle + \left\| \frac{-1}{r'_\alpha} \sum_{i=1}^{\epsilon} w_i \mathbf{u}_i \right\|^2 . \tag{8.12}$$

To simplify the second term, note from equation 4.6 that $< \mathbf{t}, \mathbf{u}_i >= w_i$. To simplify the third term, note from equation 4.8 that $\mathbf{u}_i$ are orthonormal. Q_r

therefore simplifies to

$$Q_r \;=\; \|\mathbf{t}\|^2 + \frac{-2}{r'_\alpha}\sum_{i=1}^{\epsilon} w_i^2 + \frac{1}{(r'_\alpha)^2}\sum_{i=1}^{\epsilon} w_i^2 \qquad (8.13)$$

$$=\; \|\mathbf{t}\|^2 - \left(\frac{2}{r'_\alpha} - \frac{1}{(r'_\alpha)^2}\right)\sum_{i=1}^{\epsilon} w_i^2 \left\{ \begin{array}{ll} r'_\alpha = 1 & \text{for } r_\alpha \le 1 \\ r'_\alpha = r_\alpha & \text{for } r_\alpha > 1. \end{array} \right. \qquad (8.14)$$

To understand this result, again consider Q_r for the two situations, when $\mathbf{c}$ is either inside or outside the control ellipsoid.

Firstly, for the case when $\mathbf{c}$ is inside the control ellipsoid, $r'_\alpha = 1$, which leads to

$$Q_r = \|\mathbf{t}\|^2 - \sum_{i=1}^{\epsilon} w_i^2 \qquad \text{for } r_\alpha \le 1. \qquad (8.15)$$

This result is expected because when $r_\alpha \le 1$, the remainder image is exactly the null space image. Therefore, Q_r is the square of the length of the null space image, which was given in equation 7.6.

Next consider the case when $\mathbf{c}$ is outside the control ellipsoid, i.e.

$$r'_\alpha = r_\alpha > 1. \qquad (8.16)$$

For this situation the expression enclosed in brackets in equation 8.14 is constrained to

$$0 < \left(\frac{2}{r_\alpha} - \frac{1}{(r_\alpha)^2}\right) \le 1. \qquad (8.17)$$

Thus, the length of the remainder image will be longer than the null space image, but shorter than the test image itself. This is in keeping with the geometry illustrated in the right hand diagram of figure 8.1.

8.3.2 Optimisation of ϵ and α

The Q-metric expressed in equation 8.14 has two variables that can be optimised: ϵ, the effective rank of the normal training set, and α, the width of the control ellipsoid. Hence, Q_r depends on both the choice of effective rank and the choice of the width of the control ellipsoid.

Appropriate choices for the effective rank of the cluster of normal training images was investigated in section 6.3, but a range of possible values remained.

The full range ϵ can be from 1 to p, although the Guttman-Kaiser criterion suggests that only 14–19 eigenimages are statistically significant from the training set of 77 normal radiographs (section 6.3.3). It is hoped the optimum value of ϵ might be in this range because this would confirm that the Guttman-Kaiser criterion is useful in determining the number of eigenimages to retain.

The Q-metric also depends on the width of the control ellipsoid, which is determined by α. The graph in figure 7.8 shows the location of column space images relative to an $\alpha = 0.95$ control ellipsoid. It is unknown what width of the control ellipsoid maximises the predictive value of Q_r. Noting that the abnormal images tend to be further from the centre of eigenspace, it seems likely that the most predictive width will be when normal images tend to be within the control ellipsoid but abnormal images tend to be outside. The range of possible values for control ellipsoid width is $0 < \alpha < 1$. If $\alpha = 0$, then the control ellipsoid is of zero width and r_α is undefined. If $\alpha = 1$, then the control ellipsoid needs to span all possible normal images and is therefore of infinite width.

To optimise the values of α and ϵ, the Q-metric of each test image was calculated for a range of values of α and ϵ. This was able to be performed reasonably quickly, since the limiting steps when calculating Q_r using equation 8.14 are finding $\|\mathbf{t}\|$ and $w_i \ldots w_p$. These variables do not depend on α or ϵ and so only need to be calculated once per test image, rather than for every combination of α and ϵ.

For the two experiments (presented in sections 8.3.3 and 8.3.4), α ranged from 0.01 to 0.99, with increments of 0.01, while ϵ ranged from 1 to 77. To compare the success rates of classification using different combinations of α and ϵ, the area under their ROC curves (AUC) was compared.

The use of AUC for comparing classification success was discussed in section 7.3.2. It is worth remembering that area under a ROC curve is only one way to estimate classification accuracy. For some clinical situations it may be preferable to maximise sensitivity, specificity, or another indicator of classification accuracy. For example, with a junior clinician working in the emergency department it would be sensible to maximise sensitivity, rather than AUC. This would to ensure that the clinician does not miss an abnormality on the radiograph and inadvertently send the patient home.

It should also be remembered that neither the small nor large test set is

composed of random images. Rather, the sets contain selected images to allow development of eigenimage processing. Ideally α and ϵ would be optimised using a very large set of random images.

In summary, classification was performed for a range of values of α and ϵ. The values which maximise AUC are the optimum effective rank and width of the control ellipsoid.

8.3.3   Small test set: 15 abnormal images

In this experiment, Q_r was calculated for the 15 normal and 15 abnormal test images using a range of values of α and ϵ. The test sets were those used in chapter 7, allowing direct comparison with the previous classification systems that used either the column space image or null space images alone.

Figure 8.2 shows a contour plot displaying the AUC as a function of both α and ϵ. It demonstrates that the maximum AUC is 0.800, which occurs when $\alpha = 0.05$ and $\epsilon = 15$. This is better than the AUCs for either the null space classification or column space classification systems, which had AUCs of 0.79 and 0.74 respectively. It is also worth noting that there is a "low ridge" where the AUCs are nearly identical. This ridge extends from the bottom near $\epsilon = 10$ on the graph.

8.3.4   Large test set: 74 abnormal images

The same process for optimising α and ϵ was then applied to the larger test set containing 15 normal images and 74 abnormal images. This test set was introduced in chapter 5 for studying the effect of registration. The contour plot from this experiment is shown in figure 8.3. The optimum values were similar to those of the small test set, $\alpha = 0.17$ and $\epsilon = 8$, producing a maximum AUC of 0.833. Again, there is a low ridge in the AUC along $\alpha = 0.25$ and $\epsilon = 8$ to $\alpha = 0.01$ and $\epsilon = 15$. Hence the results are in keeping with the results from the small test set.

The maximum AUC is higher than that from the smaller test set, implying that the test images used for the larger set are easier to differentiate from the normal images than those in the small test set. This finding, coupled with the maximum AUC occurring along a low ridge, suggests that optimisation should
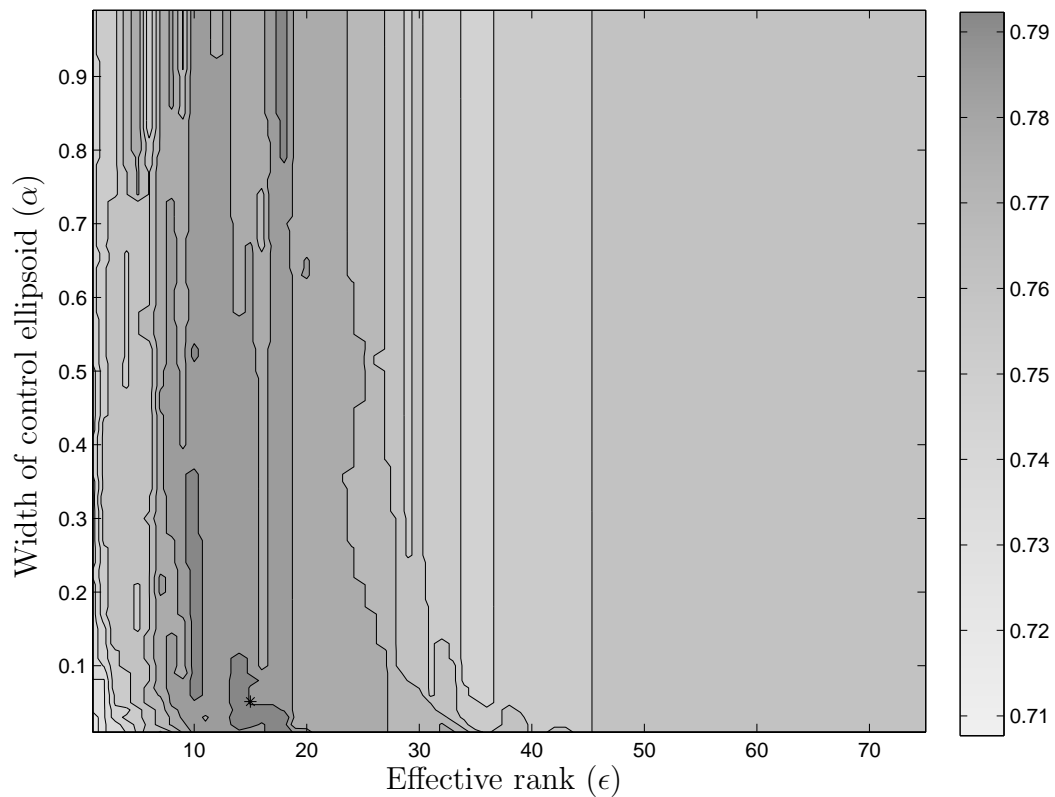
Figure 8.2: **Optimisation of α and ϵ, small test set:** This contour plot shows the relative values of AUC as a function of α and ϵ. Darker areas represent higher values of AUC. The asterix marks the maximum value of AUC, 0.800, which occurs at $\alpha = 0.05$ and $\epsilon = 15$.

be performed with a very large test set containing random images.

8.3.5 Optimum effective rank

The optimum predictive value of the remainder image is when the effective rank is chosen to be in the range $\epsilon = 8$ to $\epsilon = 15$. Increasing the ϵ above this range decreases the predictive value of the remainder image. This finding is consistent with the Guttman-Kaiser criterion, which suggested that only 14–19 eigenimages are statistically significant. Again, it would be useful to have a much larger test set than is available for this thesis to more accurately identify the optimum effective rank.
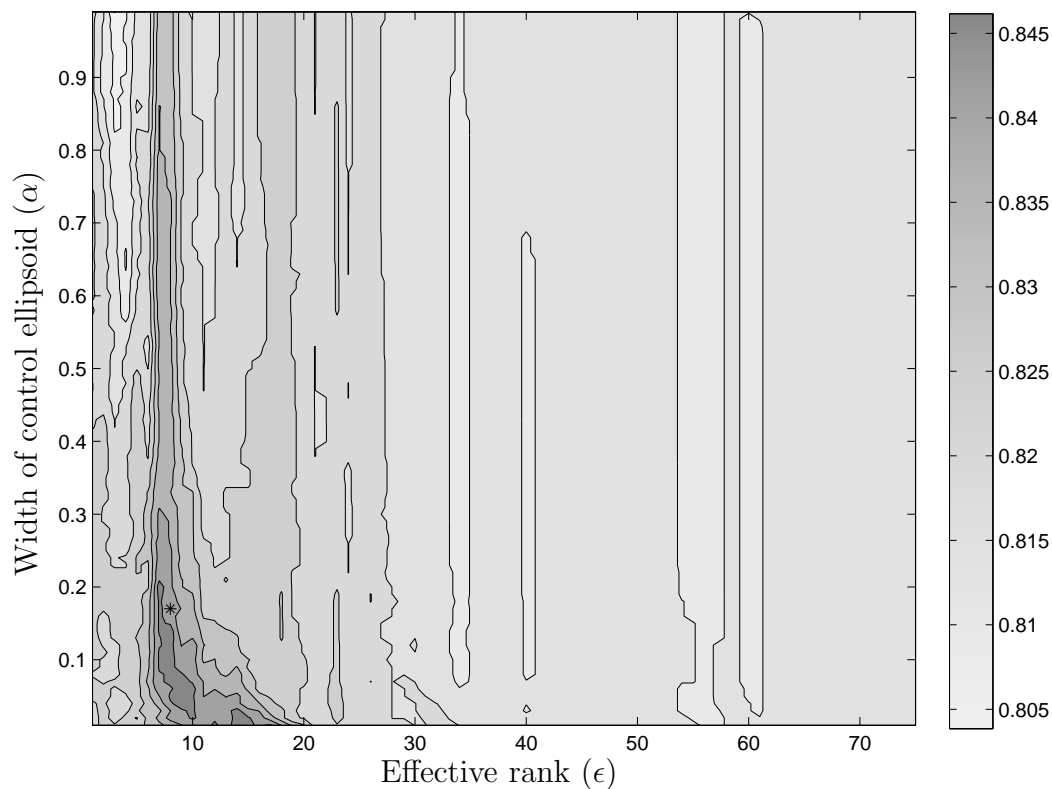
Figure 8.3: **Optimisation of α and ϵ, large test set:** This contour plot shows the relative values of AUC as a function of α and ϵ. The asterix marks the maximum value of AUC, 0.817, which occurs at $\alpha = 0.02$ and $\epsilon = 14$. There is a low ridge of high AUC values lying in the region of $\alpha = 0.25$, $\epsilon = 8$ to $\alpha = 0.01$, $\epsilon = 15$. The maximum value lies off this ridge due to small variations in the surface. A more sophisticated approach would be to smooth the surface before choosing the maximum value of AUC.

The Guttman-Kaiser criterion attempts to determine which eigenimages to retain by testing for statistical significance. It assumes that non-statistically significant eigenimages do not provide a good basis for representation of the variance within the normal set.

8.3.6 Optimum ellipsoid width

The optimisation of $\alpha \approx 0.10$ means that the control ellipsoid spans about 10% of the expected range of normal column space images. Readers should note that α is an indirect measure of control ellipsoid width. Specifically, it is the proportion of normal images spanned by the control ellipsoid. Figure 7.8 shows that column space images closest to the centre are about half way between the centre and the edge of the 95% control ellipsoid. This implies that for $\alpha = 0.02$, the width of the control ellipsoid is around one third that of an $\alpha = 0.95$ control ellipsoid.

To understand the results of optimising α, cumulative probability distributions were plotted for the location of the column space images relative to the optimum control ellipsoid ($\alpha = 0.17$ and $\epsilon = 8$). These distributions are shown in figure 8.4 and demonstrate that for the more normal images (i.e. small values of r_α), $r_{0.17}$ is approximately one or less, implying that the entire column space image is removed from the test image. For the more abnormal images (i.e., large values of r_α) $r_{0.17}$ is in the range of 1.5 to 3. Thus, for these more abnormal images approximately 30–60% of the column space image is removed from the test image.

Figure 8.4 demonstrates that the control ellipsoid is at optimum width when most normal, and all pathological images lie outside it. Thus the conjecture is: the optimum width of the control ellipsoid is when as much of the column space image is removed as possible, while ensuring that the portion of the column space image not removed can discriminate between normal and abnormal.

8.3.7 Classification at optimum ϵ and α

Having found values for the ϵ and α that maximise AUC, it is interesting to return to the Q_r values that produced this AUC. The distributions of the Q_r metric for normal and abnormal images are shown in figure 8.5. The distributions are more separated than the Q-metric distributions in the last chapter which used the null
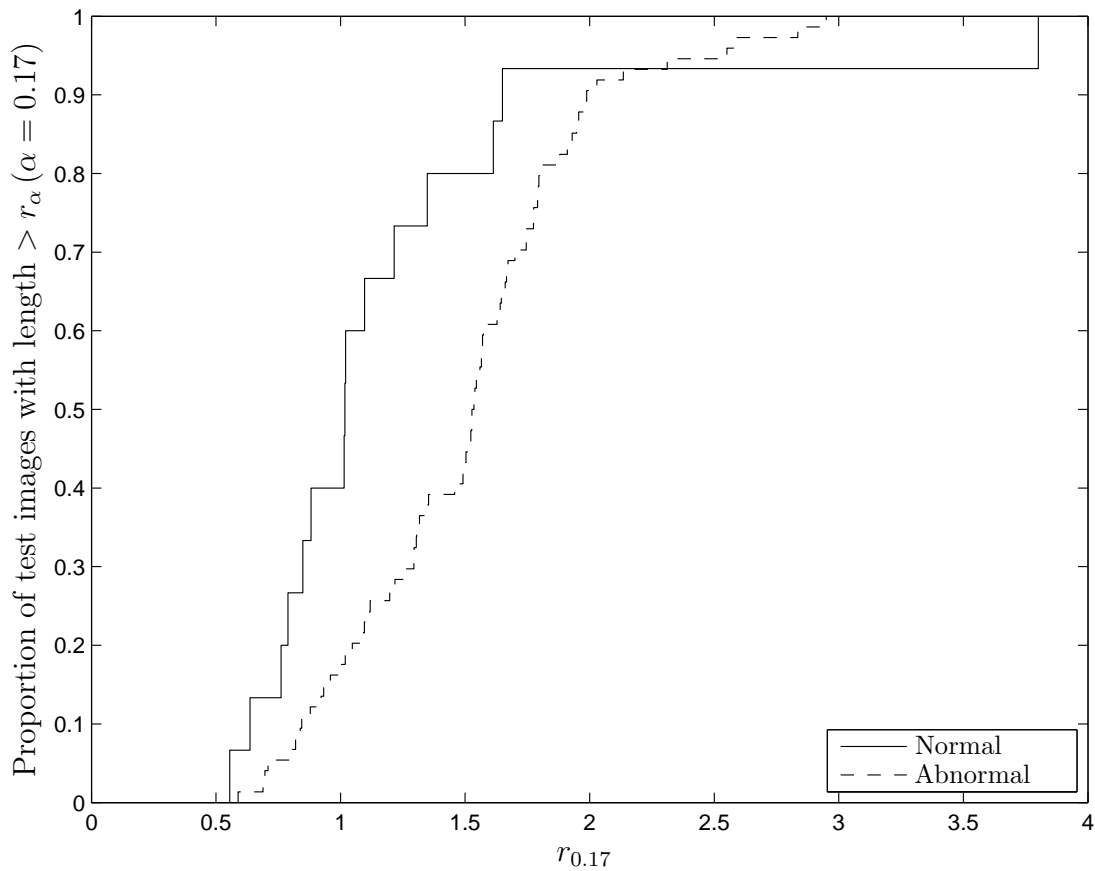
Figure 8.4: **Location of image on the optimum control ellip-soid:** In this plot the edge of the control ellipsoid is at $r_{0.17} = 1$. This corresponds to the **g** in figure 8.1.

space images alone, figure 7.12.

Using the Q-metric at $\alpha = 0.17$ and $\epsilon = 8$ it is possible to produce a ROC curve. Specifically, this is the ROC curve with the maximum AUC, and is shown in figure 8.6. It is a better ROC curve than those from the column space and null space classification schemes, figures 7.9 and 7.13 respectively.

8.4 Scattergram of test images

The location of the test image along the axis representing the ith eigenimage is given by w_i. In addition, the length of the null space component can be established from equation 7.6. Thus it is possible to produce scattergrams of
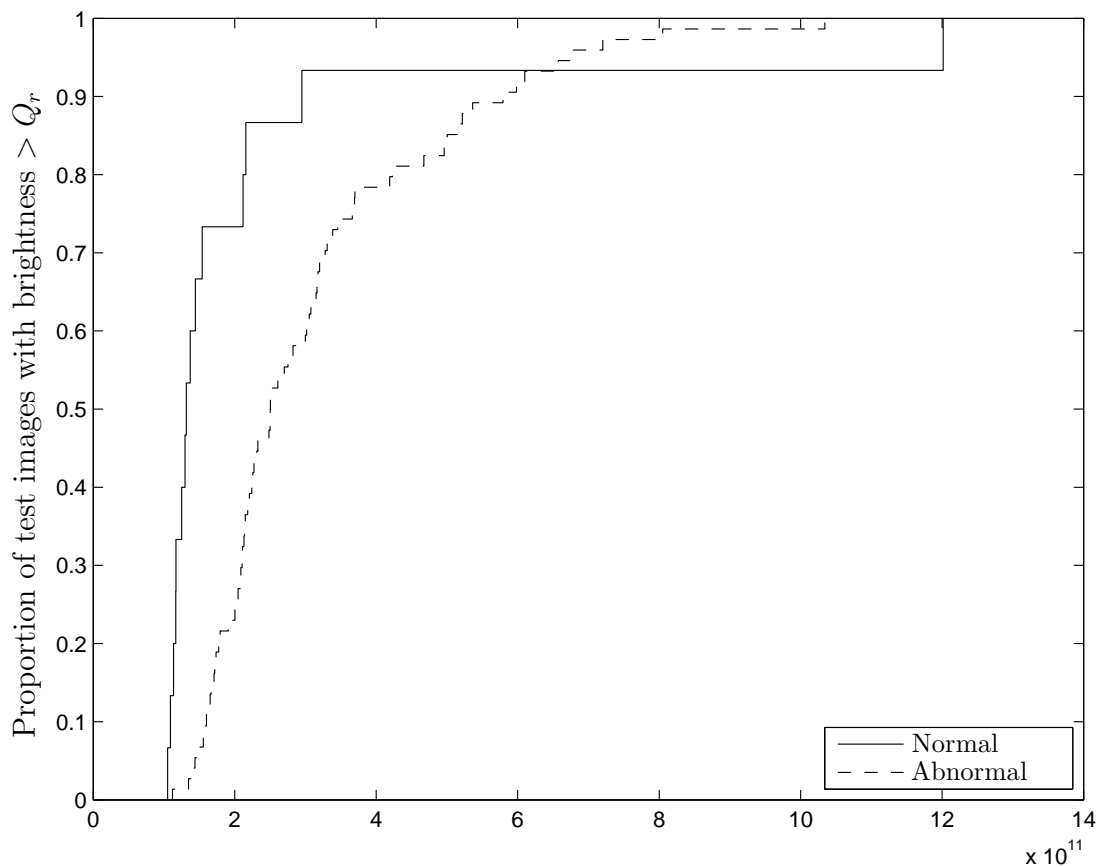
Figure 8.5: **Values of the Q-metric at optimum ϵ and α:** Using $\alpha = 0.17$ and $\epsilon = 8$ the cumulative probability distributions of Q_r are plotted for the test set containing 15 normal and 74 abnormal images.

real data, presented in the form used in figure 8.1. A scattergram for the 15 normal chest radiographs is presented in figure 8.7. The optimum control ellipsoid ($\alpha = 0.17$ and $\epsilon = 8$) is shown on the plane formed by the first two eigenimages.

It should be noted that in the plane of the first two eigenimages, most column space images lie within the control ellipsoid even though the graph in figure 8.4 demonstrates that for maximum predictability most column space images lie outside the control ellipsoid. The explanation for this is that for a column space image to lie outside the control ellipsoid, it only needs to be outside it in one of the $\epsilon = 8$ dimensions.
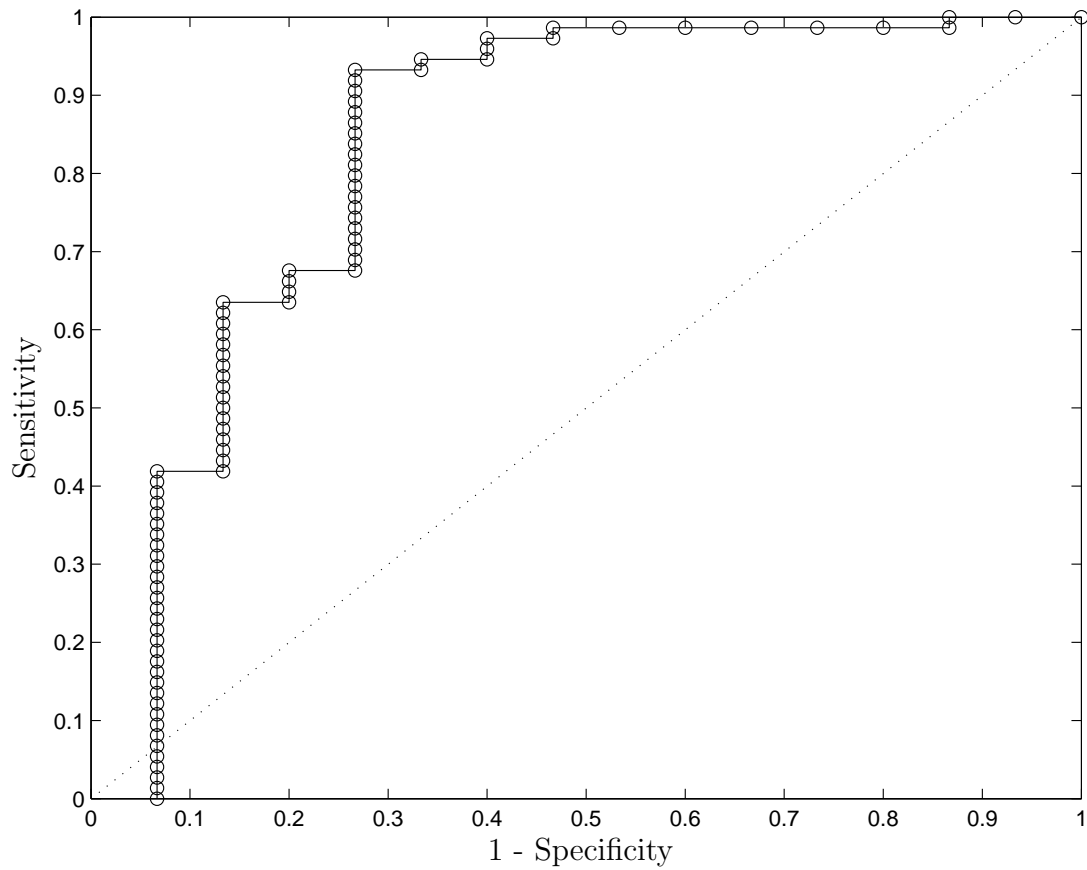
Figure 8.6: **Optimum Receiver Operating Characteristic curve for classification using the remainder image:** Using $\alpha = 0.17$ and $\epsilon = 8$, image classification was performed to assign test images as either normal or abnormal. These values of α and ϵ produce the optimum ROC curve, with an AUC of 0.837.
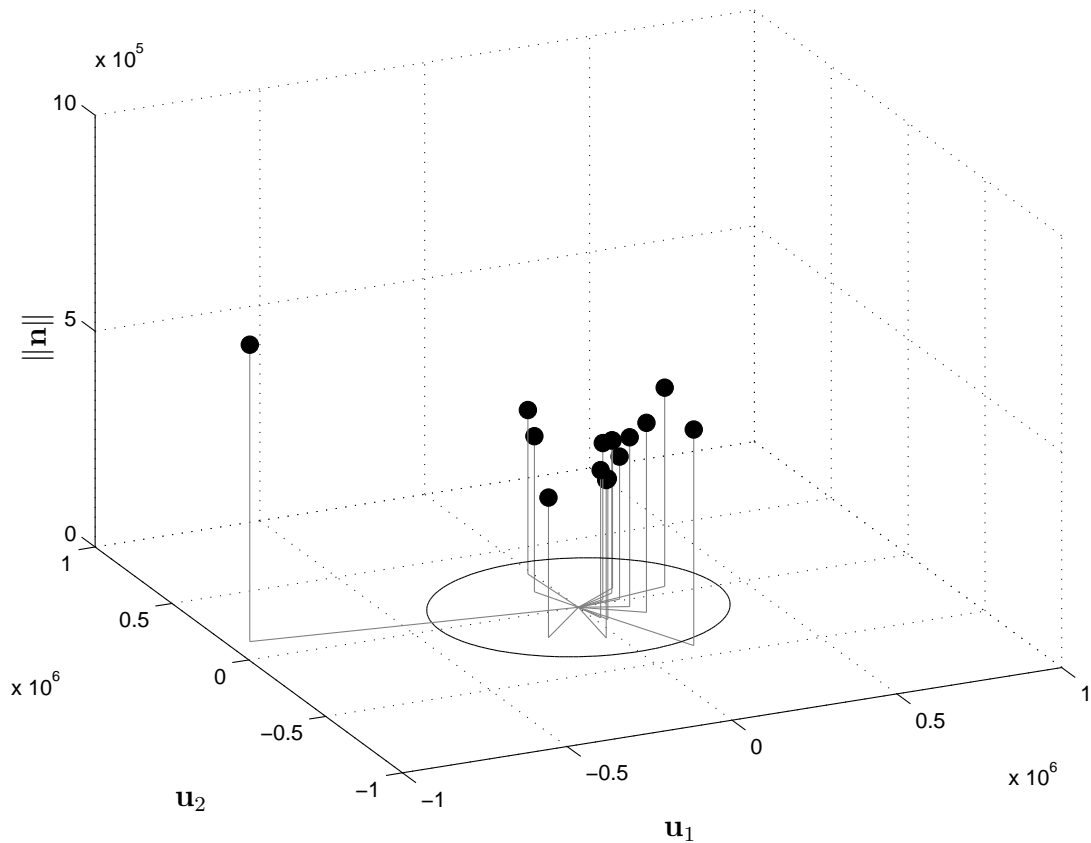
Figure 8.7: **3D scattergram of normal images:** This scattergram mimics the diagram in figure 8.1. Each dot represents one test image. The x-y plane is the first two dimensions of the $\epsilon = 8$ dimensional control ellipsoid. The width of the ellipse plotted on this plane is the ellipsoid with the most predictive value, i.e. $\alpha = 0.17$. The z-axis plots the length of the null space image. The null space is a very high dimensional space, with basis $\mathbf{u}_{\epsilon+1} \ldots \mathbf{u}_k$, and is therefore very difficult to represent in a diagram. There is one outlier, and this is discussed in the text.

8.5 Viewing the remainder image: enhancing abnormal features

The remainder image is what is left when patterns from the training set of normal images are removed. These remainder images can be used as a form of image enhancement, similar in many ways to other enhancement systems, such as temporal subtraction [12, 14, 15]. In temporal subtraction, the enhanced image is the difference between a current image and a previous image, making areas of new pathology more apparent. Similarly, the remainder image demonstrates the difference between the actual image and what is represented by the training set of normal images. It is possible that the removal of the "normal" component will make pathology more apparent to the clinician.

Because the remainder image is a "difference image," it has been shown as an absolute image, similar to the null space images of section 7.4.1. Black pixels are where the pixel intensities in the source image are well explained by the control ellipsoid formed from the normal training set. Conversely bright (white) pixels are where the pixel intensities in the source image are not well explained by the control ellipsoid. For the purpose of this thesis, remainder images shown with these display parameters will be referred to as enhanced images.

For each test image in the following sections both the original image and the remainder (enhanced) image are presented. The original image is a radiograph after registration, but before intensity centring. Specifically, it is the test image plus the average image: $\mathbf{a} + \mathrm{E}\left[\mathbf{a}\right]$.

To allow comparison between images the window and level settings (intensity range from black to white) are the same for all original and enhanced images.

8.5.1 Normal images

Figure 8.8 shows enhancement of a normal chest radiograph. This is a normal test image, rather than a normal image from the training set. Although there is some brightness in the enhanced image, it is generally quite dark. In other words, most of the patterns found within this normal test image are also found within the normal training set.

<div align="center">Original image   Enhanced image</div>
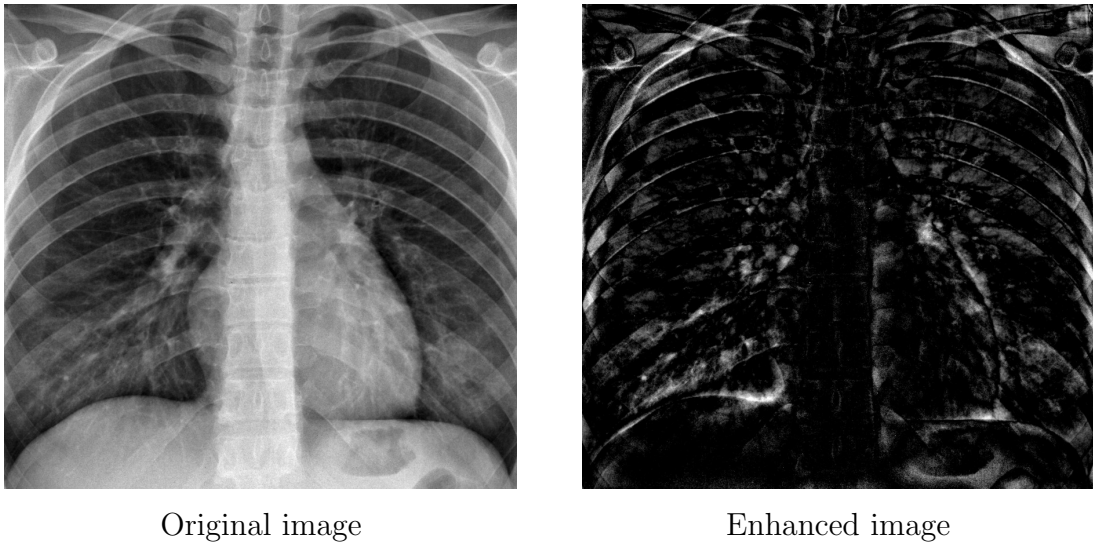
Figure 8.8: **Enhancement of a normal test image:** A normal test image shown before and after enhancement. The image on the left is the image following registration. The image on the right is the remainder image.

8.5.2 Abnormal images

Figure 8.9 shows an example of right upper lobe pneumonia. In the enhanced image the area with pneumonia is brighter than the normal areas. This is likely to be because the area of abnormality in the right upper lobe is not within the control ellipsoid of the normal training set. The successful enhancement of this pathology is a pleasing result.

Figure 8.10 demonstrates enhancement of a chest mass, in this case a probable carcinoma. The area of abnormality is bright on the enhanced image, similar to the right upper lobe pneumonia. The left upper corner of the enhanced image (i.e., the patient's right side) is also very bright, because the chest mass has affected the shape of the chest cavity. The deformed chest cavity made image registration difficult for this corner of the image, producing an area not well represented by the normal training set.

Figure 8.11 demonstrates a chest drain in a pneumothorax. The chest drain is not well seen in the enhanced image. This is probably because it covers a
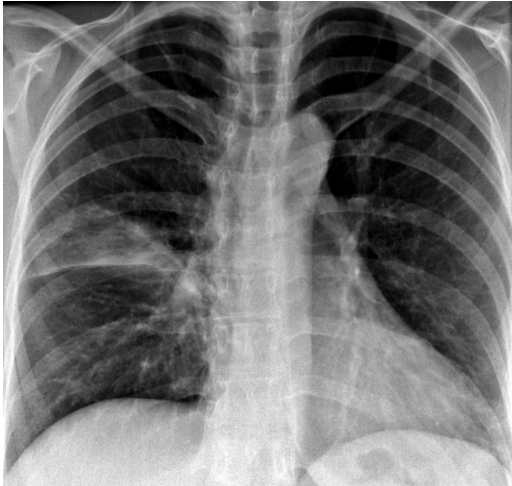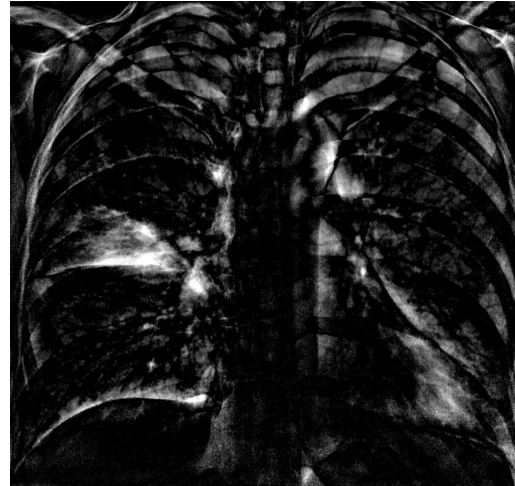
Original image   Enhanced image

Figure 8.9: **Enhancement of a right upper lobe pneumonia:** The area of abnormality in the source image is very bright in the enhanced image. It is hoped that this will improve a clinician's ability to identify the abnormality.

small area of the image, is not very bright, and mimics normal ribs. The nipple jewellery is well seen because the metal appears as a very bright structure in the source image. The pneumothorax enhances strongly despite the low intensity of its change in the original image. This is likely to be because it is large and therefore present in many pixels.

8.6   An outlier image

In figure 8.5 it can be seen that there is one test image that has a Q-metric about 3 times that of any other normal image. In fact this normal image has a Q-metric greater than any of the truly abnormal images. This results in the ROC curve in figure 8.6 crossing the diagonal line. This image also shows up as an outlier on the scattergram in figure 8.7, as well the probability distributions plotted in figures 7.8, 7.12, 8.4.

The original and enhanced forms of this outlier image are shown in figure 8.12. It can be seen that the original image is under-exposed, and this has resulted

Original image   Enhanced image
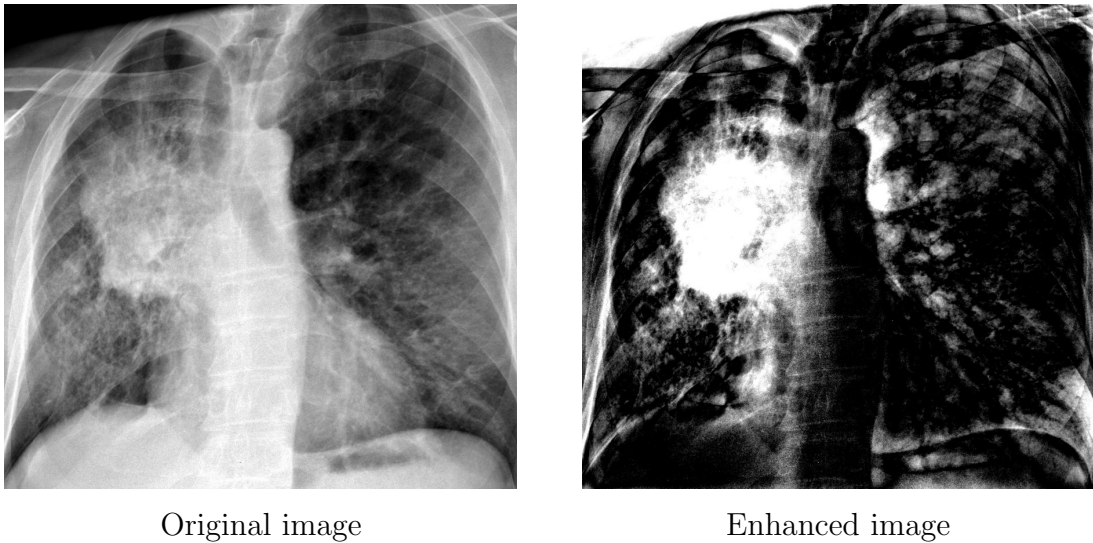
Figure 8.10: **Enhancement of a right central chest mass:** The large central chest mass has enhanced well. The left upper corner of the enhanced image is also bright because of the deformed chest cavity.

in the image being too bright. This was not appreciated early in the research because some of the viewing software automatic determination of window and level values. In particular, automatically determining window and level settings is MATLAB's default behaviour. When viewing the image with the same window and level settings as other images (in figure 8.12) the under-exposure that caused this image to be an outlier becomes more apparent.

The registration system used for this thesis makes no attempt to account for exposure variations because the radiographer manually performs intensity scaling to ensure that the radiographs all appear similar using the hospital's default window and level values. A review of all the other training images and test images indicated that no other images were of an incorrect intensity. The most likely explanation for the outlier is that the radiographer incorrectly adjusted the image intensity, or forgot to do so. The reporting radiologist is likely to have simply adjusted the viewing parameters during their reporting process without mentioning it in their report.

In summary, "normal image 088" is an outlier due to technical reasons as-
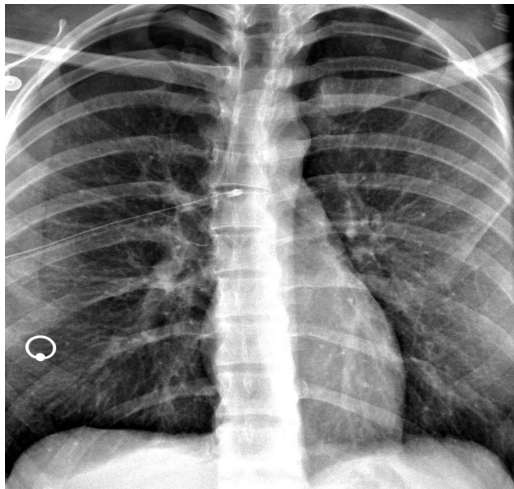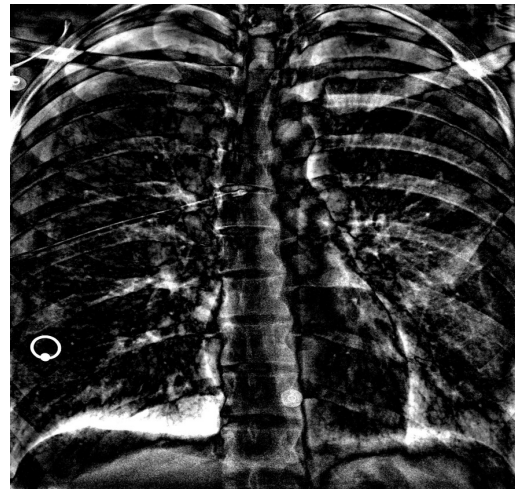
| Original image | Enhanced image |

Figure 8.11: **Enhancement of a right pneumothorax with chest drain in situ:** The bright area just above the right hemi-diaphragm is the sub-pulmonic component of the pneumothorax. It is well seen on the enhanced image, but less well seen on the original image; thus enhancement could potentially aid a clinician in spotting the abnormality. The image also contains a chest drain and nipple jewellery.

sociated with image acquisition. In hindsight a more rigorous quality assurance process for source images may have excluded the image. Possible methods include both subjective assessment and objective measurements of images [154]. However, it is pleasing that it was readily identified using the remainder image and Q-metric. In addition, it highlights the need to assess the system with a large set of random images.

The outlier image was removed from the test set and the optimisation of AUC performed again, generating the ROC curve shown in figure 8.13. Excluding the outlier image brings the system's maximum predictive score (AUC) to 0.897 with $\alpha = 0.13$ and $\epsilon = 8$, a very encouraging result.

8.7   Synthetic images

The remainder image combines information from the column space image and null space image for classification. It is possible to generate images that the clas-

Original image   Enhanced image
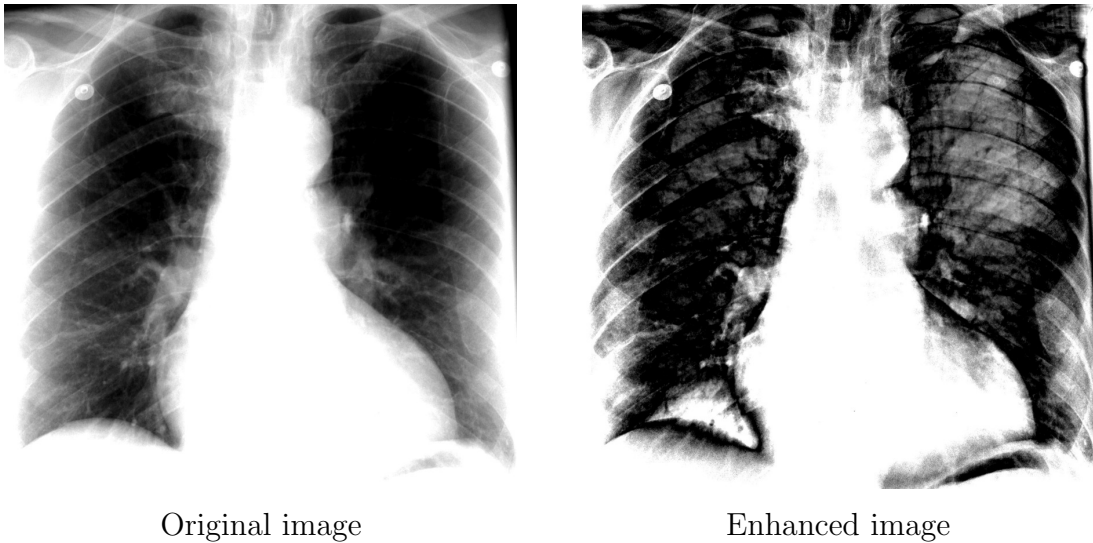
Figure 8.12: **Enhancement of the outlier image from the normal training set:** The image on the left is the original image. It is presented here using the same window and level settings that were used for all the other images. If window and level settings are chosen to be approximately twice those of other images then the radiograph appears normal. As discussed in the text, this required change in display settings is most likely due to failure of the radiographer to manually adjust the brightness. The brightness of the enhanced image thus reflects that the image is underexposed, rather than containing pathology.

sification system will consider normal, but are in fact generated using a random location within the control ellipsoid. These images are referred to as "synthetic" images. The purpose of producing synthetic images is to establish that since the control ellipsoid circumscribes normal images, choosing a random point within the control ellipsoid should visually appear similar to a normal radiograph.

One way to generate a synthetic image which will be classified as normal is to choose $\mathbf{c}$ to lie within the control ellipsoid and then choose $\mathbf{n}$ so that it is both orthogonal to $\mathbf{c}$ and of similar length to measured null space images. Choosing $\mathbf{c}$ and $\mathbf{n}$ to have these characteristics ensures that Q_r is of a similar value to normal test images.

The method for generating a synthetic normal image $\mathbf{z}$ is summarised as:
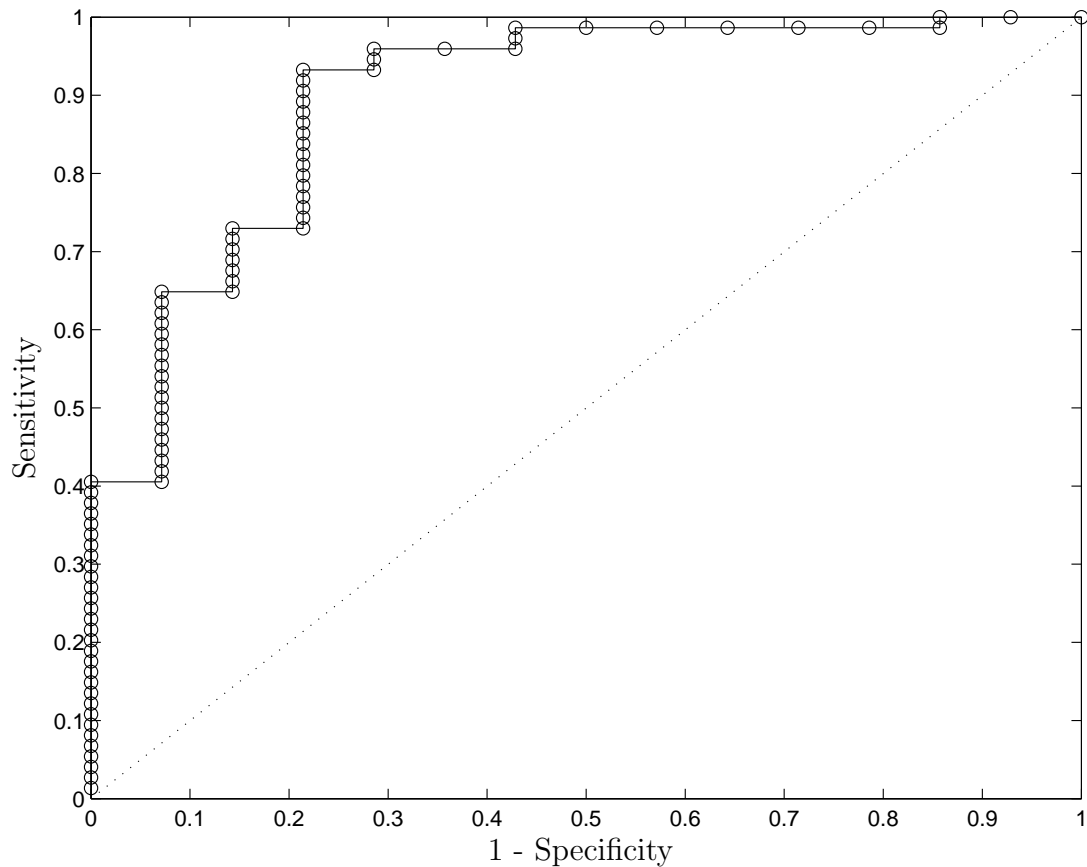
Figure 8.13: **Optimum Receiver Operating Characteristic curve for classification when the outlier image is removed:** Using $\alpha = 0.13$ and $\epsilon = 8$, image classification was performed to assign test images as either normal or abnormal. These values of α and ϵ produce the optimum ROC curve, with an AUC of 0.897.

1. Generate a column space image that lies within the control ellipsoid. Store this as $\mathbf{c}_z$.

2. Generate an arbitrary image, say $\mathbf{x}$, to use as a starting point for the null space image.

3. Use Gram-Schmidt orthogonalisation to make $\mathbf{x}$ orthogonal to $\mathbf{c}_z$. Store this as $\mathbf{y}$.

4. Multiply $\mathbf{y}$ by a scalar b, so that the length of $b\mathbf{y}$ is similar to measured null space images. Store $b\mathbf{y}$ as the null space component $\mathbf{n}_z$.

5. Add the null space and column space components together to form the synthetic image, $\mathbf{z}$.

Using this technique the column space will always be within the control ellipsoid and so the remainder image will be exactly represented by the null space image. Scaling the null space image ensures that Q_r is always of a value that is considered normal.

8.7.1 Detailed method of producing synthetic normal images

A column space image that lies within the control ellipsoid is generated by adding together the first ϵ eigenimages, with the associated weighting factors, w_i. Specifically,

$$\mathbf{c}_z = \sum_{i=1}^{\epsilon} w_i \mathbf{u}_i \qquad (8.18)$$

where each w_i is randomly chosen from a normal distribution with a mean of zero and a variance of λ_i.

A k-dimensional vector $\mathbf{x}$ with randomly chosen elements is used as the starting point for finding $\mathbf{n}_z$. This starting vector is then manipulated using Gram-Schmidt orthogonalisation to remove all parts of $\mathbf{x}$ that are in the same direction as each eigenimage [26].

$$\mathbf{y} = \mathbf{x} - \sum_{i=1}^{\epsilon} < \mathbf{x}, \mathbf{u}_i > \mathbf{u}_i \qquad (8.19)$$

While $\mathbf{y}$ now has the most important property of null space images (it is orthogonal to the column space image) it may be so long that $\mathbf{n}_z$ would be classified as abnormal. Hence, the length of the null space component must be made similar to that of normal test images. From figure 7.2 it can be seen that for normal test images the expected ratio of the length of the column space image to the length of the null space image ($\frac{\|\mathbf{n_z}\|}{\|\mathbf{c}_z\|}$) is 1.2. Therefore $\mathbf{y}$ needs to be multiplied by a scalar b so that the length of $\mathbf{n}_z$ is appropriate. More formally,

$$\mathbf{n}_z = b\mathbf{y} \tag{8.20}$$

where b is found from

$$b = 1.2\frac{\|\mathbf{c_z}\|}{\|\mathbf{y}\|} \tag{8.21}$$

Finally, the synthetic image is formed by adding its null space and column space components

$$\mathbf{z} = \mathbf{c}_z + \mathbf{n}_z. \tag{8.22}$$

8.7.2 Viewing synthetic normal images

Figure 8.14 contains a synthetic image and its enhanced version. The display parameters are the same as those used for all other original and enhanced images shown in this chapter.

The synthetic image was generated using Gaussian noise as a starting point for producing the null space image. That is, $\mathbf{x}$ was chosen such that each pixel's intensity was normally distributed with a mean of zero and variance of one. Gram-Schmidt orthogonalisation was performed to ensure that $\mathbf{y}$ was orthogonal to the column space, and then scaled to ensure that the relative length of the column space image and null space images was 1.2. The column space image is a random point within the control ellipsoid.

The synthetic image in figure 8.14 looks like a normal, but noisy, chest radiograph. The enhanced image is dark with the only features resulting from the Gaussian noise and orthogonalisation used in the image's construction. This result was expected because the synthetic image was generated to be a "normal image".

The result is that a synthetic image can be chosen to lie within the control

Original image   Enhanced image
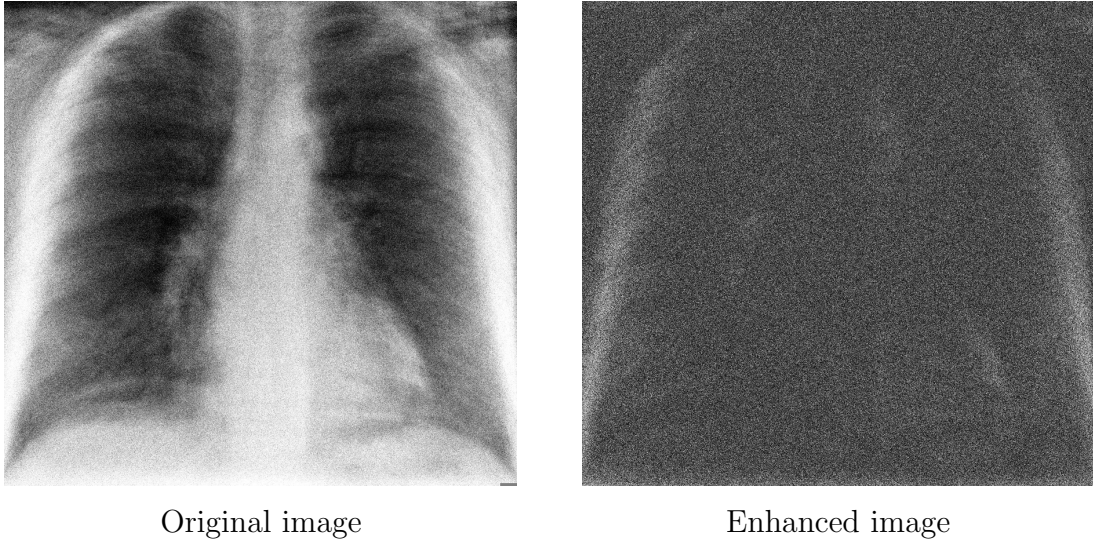
Figure 8.14: **Enhancement of a synthetic image:** The image on
the left is a synthetic normal image. It has been generated by setting
the column space component to be a random point within the cluster
of normal images, while the null space image uses Gaussian noise as
its initial value. The enhanced image is dark and featureless.

ellipsoid. Such images visually appear similar to a normal radiographs. This
is a helpful result because it adds further weight to the conjecture that normal
images are clustered within the control ellipsoid.

8.8   Conclusion

Previous chapters in this thesis concluded that applying Principal Components
Analysis (PCA) to a training set of normal radiographs allows identification of
a region within image space that represents "normal images". The eigenimages
are an orthonormal set of basis vectors for this region. The width of this region
is found using the characteristic roots. This region is referred to as the control
ellipsoid. Normal chest radiographs are expected to be clustered in this region.

Once the control ellipsoid has been found it is possible to remove parts of a
test image that lie within this volume. The part not within the volume is referred
to as the remainder image. This remainder image can be used for classification

171

schemes using the Q-metric. Alternatively, it can be displayed as an enhanced image in which areas of abnormality are brighter than normal areas.

For the test set of 15 normal images and 74 abnormal images the maximum predictive value (area under the ROC curve) was 0.837. This is a good result and confirms that using the Q-metric of a remainder image is better for classification than earlier schemes based solely on the null space or column space images. This was an expected result because the remainder image combines information regarding the distance of a test image from the normal cluster, both with respect to the distance from the hyper-plane defined by the principal components and the in-plane distance from the centre of normal.

An outlier included in the 15 normal test images is thought to result from failure of manual intensity correction by the radiographer. Excluding the outlier image from the test set improves the classification accuracy to nearly 90%. Although images used in this thesis should have had their intensity manually corrected by the radiographer, it would be possible to add an intensity correction step to the registration process if required.

Optimisation of the two parameters of Q_r, namely ϵ and r_α, suggests that for maximum predictability the dimensionality of the normal cluster is near to the number of significant principal components found using the Guttman-Kaiser criterion. The optimal width of the normal cluster occurs when normal images are just at the edge of the control ellipsoid. This maximises the separation of normal and abnormal images in the hyper-plane of the control ellipsoid.

8.9   Summary

1. The remainder image is a derived image that can be used for classification and enhancement schemes.

2. The maximum predictive value achieved was 0.897. This value was determined using the selected sets of normal and abnormal images. The maximum predictive value for a larger random set of images may be different.

3. The optimum effective rank for maximum predictive value is near to the

Guttman-Kaiser criterion for estimating the number of significant components.

4. The optimum width of the normal control ellipsoid for maximum predictive value is such that as much as possible of the normal variance is represented, while still ensuring that abnormal images lie outside the control ellipsoid.

5. Depending on the clinical situation, the registration process may need to include correction for exposure variation between radiographs.

Chapter IX

Conclusion

Eigenimage processing is a technique that identifies a cluster of images within a high dimensional image space. This cluster is described by a small number of orthogonal basis vectors, with each image's location being a limited distance along each basis vector. In particular, when using a normal training set, eigenimage processing identifies a cluster of normal images and allows the concept of "normal" to be expressed in a particular way:

- The centre of the cluster of normal images is the average chest radiograph. Variations of normal tend to be close to this.

- There is a limited number of ways in which normal images tend to vary about the average image. These common patterns of variance are the eigenimages.

- For each pattern of variance there is a limited magnitude of variation, given by the characteristic roots, which may still be considered normal.

The development of the remainder image used this description of normal to subtract normal features from test radiographs. The remainder image can be used for classification or be presented directly to a clinician as a form of image enhancement.

9.1 The work in this thesis

Chapters 1, 2, and 3 introduced the radiological problem, background information on eigenimage processing, and the mathematical method. Experiments performed in later chapters implemented and refined a classification and enhancement system suitable for the radiological problem.

175

Development of the prototype system, chapter 4, demonstrated that eigen-image processing can be applied to high resolution chest radiographs. This is significant because to date very little work has been published on the application of Principal Components Analysis (PCA) to medical images. Chest radiographs are a large 2-dimensional medical image, hence the success in building the prototype system suggests that PCA can be applied to other similarly sized medical images. This conclusion is significant to both clinical radiology and image processing.

The work on image registration, chapter 5, was important because it highlighted the complexity of alignment problems. One registration system that was ideal for all situations could not be identified. However registration using an affine transformation based on manually selected control points was found to be a reasonable compromise. It is simple, fast, and reliable in the presence of pathology. This registration system could be integrated into clinical practice with either a radiographer or radiologist identifying the landmarks. None of the fully automatic registration methods assessed were fast enough for clinical use, and were very unreliable when significant pathology was present.

Chapter 6 studied the application of PCA to a training set of normal images. It demonstrated that PCA is good at identifying and describing the cluster of normal images. The chapter also concluded that the number of significant patterns of variance will usually be much less than the number of images within the training set. The experiments concluded that a training set of 77 normal images was too small, and that 200-300 images would be better. The time to process such a training set is reasonable on modern computer hardware. The chapter also concluded that increasing the image size (e.g. higher resolution or 3-dimensional images) linearly increased the computation time for increasing numbers of pixels or voxels. This is a significant result as it allows future work on much larger images, such as those generated by CT or MRI.

The work in chapter 7 showed that both the column space and null space images contain diagnostic information. Turk and Pentland's original approach used only the column space image. This difference arises because in their work the training set represented all possible images, whereas for the work in this thesis the training set only contains normal images. Knowing that both the column space and null space images contain diagnostic information changed the

approach to the classification scheme. In particular it became clear that the system would need to include information from both component images.

Chapter 8 developed a new derived image, the remainder image, which combined information from the column space and null space images. It showed that the remainder image was better for a classification scheme than either the column space or null space image alone. The work identified the optimal width and effective rank of a control ellipsoid which circumscribes the cluster of normal radiographs. Unfortunately it was evident that a much larger set of test images should be used in the optimisation process.

For the remainder image, the Q-metric can be used as a measure of "how abnormal" an image is. Classification using the Q-metric gave a maximum predictive value of around 90%. This is very encouraging, although it must be remembered that the value was calculated using selected images. It is unknown what the value would be if the test set contained a random selection of abnormal images.

In addition to its use in a classification scheme, the remainder image can also be shown to clinicians to help identify areas of abnormality within an image. This work looks promising, but its clinical utility needs to be evaluated by a range of clinicians with differing radiological skills.

9.2   Future Work

The successful application of eigenimage processing to frontal chest radiographs is encouraging. It suggests that other related statistical image processing routines may also be helpful for medical image processing. In addition it implies that the technique could be applied to other image types such as those from CT and MRI. It is hoped that eigenimage processing will find routine use within radiology.

9.2.1   Related statistical techniques

One limitation of PCA is that it seeks an efficient basis to represent variance, but makes no attempt to find a basis that maximises the ability to differentiate between classes. The derivation of the remainder image takes advantage of this by using PCA to efficiently represent normal images. However, it would be interesting to try other statistical techniques, such as Linear Discriminant Anal-

ysis (LDA), that aim to maximise discrimination [56]. Another possibility is to combine two or more statistical techniques into a single system [51].

Another limitation of PCA is that it identifies patterns of linearly independent variance. It would be useful to attempt non-linear analysis. An example of linear variance of two correlated variables was given in figure 2.1. If the graph were to have the observations scattered around a circle then PCA would struggle to identify two linear patterns of variance, even though the variable would clearly not be independent. Jackson's text discusses this in detail [25]. Methods for non-linear component analysis include Kernel-PCA, curvilinear components analysis, and independent components analysis [44, 51, 54].

9.2.2 Increasing the number of images

The study presented in this thesis used an arbitrary number of images for both the training and test sets. Experiments in chapter 6 suggested that a larger training set could be helpful in identifying more significant eigenimages. Experiments in chapter 8 suggested that a larger test set could improve the optimisation process and better establish the strengths and weakness of the system.

Using the Guttman-Kaiser criterion the training set of 77 normal images produced 14–19 significant eigenimages. This number was in keeping with the optimal rank of the control ellipsoid. It is hoped that by extending the training set to several hundred images many more significant eigenimages could be identified. This should improve the classification scheme and image enhancement.

With a view to expanding or incrementally updating the training set, there has been significant work on combining eigenspaces; the most promising work is by Hall *et al.* [144–146]. Using their method it would be possible to have a training set in the order of hundreds to thousands of images.

In chapter 8, there were difficulties in identifying optimal values for the effective rank and width of the control ellipsoid. While the experiments identified a range of possible values, a larger test set would allow for better optimisation. In addition the images in the test set were not randomly selected and to better assess the system a random selection of test images should be used.

9.2.3 Extension of the work to other image types

Within radiology, chest radiographs are often quickly interpreted by radiologists, while the large data sets from CT and MRI can be very slow to interpret. For these large data sets a significant amount of time is spent identifying abnormal features within hundreds or thousands of normal images. If eigenimage processing could be applied to CT or MRI images then its impact on clinical practice may be greater because the potential time saving is greater. There are three main issues to be explored when extending to these larger data sets: registration, training set size, and computer processing time.

Image registration problems are very different for 3-dimensional images compared with the 2-dimensional case. In some respects registration can be easier since there are no overlapping features, but the complexity of the image can make it harder. Fortunately, as noted in chapter 5, 3-dimensional image registration is a well researched area.

An important issue in extending the technique to other image types is determining the size of the training set required. Large complex 3-dimensional data sets may require a very large number of training images to produce enough statistically significant eigenimages. Again, this is an area where incremental PCA techniques may be helpful [144–146].

Chapter 6 showed that the processing time for PCA tends to be linear when increasing the image size. Images from CT and MRI can be up to 10–20 times as large as chest radiographs (in term of pixels). While this means processing would be slow on current hardware, it is likely to be practical within the near future.

9.2.4 Possible uses within radiology

The enhancement of each test image is fast compared to processing of the training set. In the experiments performed processing each test image took about one minute using non-optimised MATLAB code on standard PC hardware. However, an optimised system taking advantage of the parallel nature of the computations and run on typical radiology workstations could be up to twenty times faster. This is an acceptable processing time for integration into hospital workflow. There are two obvious locations for implementing eigenimage processing within

a department's infrastructure: either as a preprocessing step before the image is sent to a radiologist's workstation, or directly in the radiologist's viewing software. Previously there have been successful attempts to integrate other forms of image enhancement into radiology department workflows as a preprocessing step [155].

Within the clinical environment the Q-metric could be used to assign a priority to each image. Currently in radiology departments there is often a worklist of unreported radiographs. The radiologist selects an image to report without prior knowledge of the possibility of it being abnormal. The Q-metric could be used to indicate how likely an image is to contain an abnormality. This would improve workflow by ensuring that the most abnormal radiographs are reported first, enabling any changes in patient management to occur as soon as possible.

Also within a clinical environment, the enhanced images could be shown alongside (or after) the original radiograph, as a form of "second check" for the reporting clinician. Such a system would be most valuable if the system was extended to larger image types. If eigenimage processing was able to alert the radiologist to the likely areas of abnormality, it could significantly improve the speed and accuracy of reporting.

Within the research field, it may be possible to use eigenimage processing to identify areas of abnormality not currently recognised by radiologists. This may be possible because the PCA often identifies patterns of variance that differ from those identified by a radiologist. In addition PCA can identify very low power patterns that may be overlooked by a radiologist.

As introduced in section 6.5.3 an important question is which normal variants should be included within the training set. This is a complex question that could only be fully assessed by a clinical trial in which several training sets are used, and the enhanced images assessed by groups of clinicians with different skill levels. The difficulty in deciding which normal variants to include in the training set arises because a skilled clinician may benefit from having normal variants brought to their attention, while a less skilled clinician may be misled into attributing too much significance to a minor finding.

In this thesis ROC curves were used to display the success of classification, using the AUC to compare two ROC curves. Although this is a common method of comparing tests within medicine, there are other methods which could be in-

vestigated [85, 86, 138]. Most importantly, the method chosen should reflect the clinical application. For example, if the goal is to assist a junior clinician to confirm an image is truly normal, then the most appropriate measure of the system's success is its ability to detect abnormality (the test's sensitivity). However, if the goal is to assist a senior radiologist to interpret a complex radiograph, then the most appropriate measure of success is the ability to identify a particular disease with certainty (the test's specificity).

9.3 Summary

1. Eigenimage processing can be successfully applied to the high resolution images found within radiology.

2. The application of PCA to a training set of images identifies the important patterns and magnitudes of variance within a training set. If the set contains normal images, then eigenimage processing identifies normal patterns of variance, with a known level of accuracy.

3. Identifying the normal patterns of variance allows these patterns to be removed from test images, leaving a remainder image. The remainder image can be used for classification or as a form of image enhancement.

4. The success of the application of PCA to radiological images suggests that other related statistical techniques could also be attempted.

5. Important future work could involve extending the technique to new image types, such as CT or MRI images. In these data sets the radiologist often has to identify abnormal features within a very large number of normal images, which can be very time consuming. Application of eigenimage processing may significantly improve the speed of image interpretation.

References

[1] J Eng, W K Mysko, G E R Weller, R Renard J N Gitlin, D A Bluemke, D Magid, G D Kelen, and W W Scott. Interpretation of emergency department radiographs: A comparison of emergency medicine physicians with radiologists, residents with faculty, and film with digital display. *American Journal of Roentgenology*, 175:1233–1238, 2000.

[2] E J Potchen, T G Cooper, A E Sierra, G R Aben, M J Potchen, M G Potter, and J E Siebert. Measuring performance in chest radiography. *Radiology*, 217(217):456–459, 2000.

[3] Y Jiang, R Nishikawa, R Schmidt, A Toledano, and K Doi. Potential of computer-aided diagnosis to reduce variability in radiologists' interpretations of mammograms depicting microcalcifications. *Radiology*, 220:787–794, 2001.

[4] H Yoshida. Multiresolution non-rigid image registration method and its application to removal of normal anatomy. In *International Conference on Image Processing*, volume 3, pages 440–445, 1999.

[5] M Turk and A Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.

[6] M A Turk and A P Pentland. Face recognition using eigenfaces. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 586–91, 1991.

[7] H Becker, W Nettleton, P Meyers, J Sweeney, and C Nice Jr. Digital computer determination of a medical diagnostic index directly from chest X-ray images. *IEEE Transactions in Biomedical Engineering*, BME-11(67-72), 1964.

[8] P Meyers, C Nice Jr., H Becker, W Nettleton, J Sweeney, and G Meckstroth. Automated computer analysis of radiographic images. *Radiology*, 83:1029–34, 1964.

[9] B Van Ginneken, B M Ter Haar Romeny, and M A Viergever. Computer-aided diagnosis in chest radiography: a survey. *IEEE Transactions on Medical Imaging*, 20(12):1228–1241, 2001.

[10] E B Cole, E D Pisano, E O Kistner, K E Muller, M E Brown, S A Feig, R A Jong, D A A Maidment, M J Staiger, C M Kuzmiak, R I Freimanis, N Lesko, E L Rosen, R Walsh, M Williford, and M P Braeuning. Diagnostic accuracy of digital mammography in patients with dense breasts who underwent problem-solving mammography: Effects of image processing and lesion type. *Radiology*, 226(1):153–160, 2003.

[11] E D Pisano, E B Cole, B M Hemminger, M J Yaffe, S R Aylward, A D A Maidment, R E Johnston, M B Williams, L T Niklason, E F Conant, L L Fajardo, D B Kopans, M E Brown, and S M Pizer. Image processing algorithms for digital mammography: A pictorial essay. *Radiographics*, 20(5):1479–1491, 2000.

[12] S Hariharan, A K Ray, and M K Ghosh. An algorithm for the enhancement of chest X-ray images of tuberculosis patients. In *Proceedings of IEEE International Conference on Industrial Technology*, volume 2, pages 107–112, Januaray 2000.

[13] D Tsai, Yongbum Lee, M Sekiya, S Sakaguchi, and I Yamada. A method of medical image enhancement using wavelet analysis. In *6th International Conference on Signal Processing*, pages 723–726, August 2002.

[14] M Tsubamoto, T Johkah, T Kazuka, N Tamiyama, S Hamada, O Honda, N Mihara, M Koyama, M Maeda, H Nokamura, and K Fujiwara. Temporal subtraction for the detection of hazy pulmonary opacities on chest radiography. *American Journal of Roentgenology*, 179:467–471, 2002.

[15] T Johkoh, T Kozuka, N Tomiyama, S Hamada, O Honda, N Mihara, M Tsubamoto, M Maeda, H Nakamura, H Sald, and K Fujiwara. Temporal subtraction for detection of solitary pulmonary nodules on chest radiographs: Evaluation of a commercially available computer-aided diagnosis system. *Radiology*, 223:806–811, 2002.

[16] H Yoshida and B Keserci. Bayesian wavelet snake for computer-aided diagnosis of lung nodules. *Integrated Computer-Aided Engineering*, 7(3):253–69, 2000.

[17] L Monnier-Cholley, H MacMahon, S Katsuragawa, J Morishita, T Ishida, and K Doi. Computer-aided diagnosis for detection of interstitial opacities on chest radiographs. *American Journal of Roentgenology*, 171:1651–1656, 1998.

[18] H MacMahon. Improvement in detection of pulmonary nodules: Digital image processing and computer aided diagnosis. *RadioGraphics*, 20:1169–1177, 2000.

[19] H MacMahon, R Engelmann, F Behlen, K Hoffmann, T Ishida, C Roe, C Metz, and K Doi. Computer-aided diagnosis of pulmonary nodules: Results of a large-scale observer test. *Radiology*, 213:723–726, 1999.

[20] B Zheng, Y Chang, and D Gur. Mass detection in digitized mammograms using two independent computer-assisted diagnosis schemes. *American Journal of Roentgenology*, 167:1421–1424, 1996.

[21] Z Huo, M Giger, C Vyborny, and C Metz. Breast cancer: Effectiveness of computer-aided diagnosis - observer study with independent database of mammograms. *Radiology*, 224:560–568, 2002.

[22] N Petrick, B Sahiner, H Chan, M Helvie, S Paquerault, and L Hadji-iski. Breast cancer detection: Evaluation of a mass-detection algorithm for computer-aided diagnosis - experience in 263 patients. *Radiology*, 224:217–224, 2002.

[23] I Jolliffe. *Principal Component Analysis*. Springer-Verlag, 2 edition, 2002.

[24] B Flury. *Common Principal Components and Related Multivariate Models*. Wiley, 2nd edition, 1988.

[25] J E Jackson. *A User's Guide to Principal Components*. Wiley-Interscience, John Wiley & Sons, Inc., 1991.

[26] Gilbert Strang. *Linear Algebra and its Applications*. Brooks Cole, 4th edition, 2006.

[27] G H Golub and C F Van Loan. *Matrix Computations*. John Hopkins University Press, 3rd edition, 1996.

[28] M Kirby and L Sirovich. Application of the Karhunen-Loève procedure for the characterization of human faces. 12:102–108, 1990.

[29] M Turk. A random walk through eigenspace. *IEICE Transactions on Information and Systems*, E84-D(12):1586–1595, December 2001.

[30] I Aravind, C Chandra, M Guruprasad, P Sarathi Dev, and R Samuel. Numerical approaches in principal component analysis for face recognition using eigenimages. In *2002 IEEE International Conference on Industrial Technology*, volume 1, pages 246–51, 2002.

[31] E Lizama, D Waldoestl, and B Nickolay. An eigenfaces-based automatic face recognition system. In *IEEE International Conference on Systems, Man, and Cybernetics 1997*, pages 174–177, 1997.

[32] N Muller and B Herbst. The use of eigenpictures for optical character recognition. In *Proceedings. Fourteenth International Conference on Pattern Recognition*, volume 2, pages 1124–1126, 1998.

[33] G I Chiou and H Jenq-Neng. Lipreading from color video. *IEEE Transactions on Image Processing*, 6(8):1192–1195, 1997.

[34] R Kuhn, P Nguyen, J Junqua, and L Goldwasser. Eigenfaces and eigenvoices: Dimensionality reduction for specialized pattern recognition. In *1998 IEEE Second Workshop on Multimedia Signal Processing*, pages 71–76, 1998.

[35] J Mi and D Huang. Image compression using principal component neural network. In *International Control, Automation, Robotics and Vision Conference*, pages 698–701, December 2004.

[36] P R Oliveira, R F Romero, L G Nonato, and J Mazucheli. Techniques for image compression: a comparative analysis. In *Proceedings of the Sixth Brazilian Symposium on Neural Networks*, pages 249–254, 2000.

[37] L Luo, M Swamy, and E Plotkin. A modified PCA algorithm for face recognition. In *IEEE CCECE 2003. Canadian Conference on Electrical and Computer Engineering*, volume 1, pages 57–60, 2003.

[38] P Watta, N Gandhi, and S S Lakshmanan. An eigenface approach for estimating driver pose. In *Proceedings of Intelligent Transportation Systems*, pages 376–381, 2000.

[39] Y Nakamura, G Fukano, H Takizawa, S Mizuno, S Yamamoto, T Matsumoto, Y Tateno, and T Iinuma. Eigen nodule: view-based recognition of lung nodule in chest X-ray CT images using subspace method. In *Proceedings of the 17th International Conference on Pattern Recognition*, volume 4, pages 681–684, August 2004.

[40] J S Taur and C W Tao. Medical image compression using principal component analysis. In *Proceedings of International Conference on Image Processing*, volume 2, pages 903–906, September 1996.

[41] B van Ginneken and B ter Haar Romeny. Automatic delineation of ribs in frontal chest radiographs. In K M Hanson, editor, *Proceedings of International Society for Optical Engineering, Medical Imaging 2000*, volume 3979, pages 852–836, June 2000.

[42] A Kouzani, F He, and K Sammut. Multiresolution eigenface-components. In *Proceedings of IEEE Region 10 Annual Conference. Speech and Image Technologies for Computing and Telecommunications*, volume 1, pages 353–356, 1997.

[43] Y Nara, J Yang, and Y Suematsu. Face recognition using improved principal component analysis. In *Proceedings of 2003 International Symposium on Micromechatronics and Human Science*, pages 77–82, October 2003.

[44] K Sohara and M Kotani. Application of kernel principal components analysis to pattern recognition. In *Proceedings of the 41st SICE Annual Conference*, volume 2, pages 750–752, August 2002.

[45] S Shan, W Gao, X Chen, and J Ma. Novel face recognition based on individual eigensubspaces. In *5th International Conference on Signal Processing Proceedings*, volume 3, pages 1522–1525, 2000.

[46] R Chellappa, C L Wilson, and S Sirohey. Human and machine recognition of faces: a survey. In *Proceedings of the IEEE*, volume 83, pages 705–741, May 1995.

[47] B Moghaddam. Principal manifolds and Bayesian subspaces for visual recognition. In *The Proceedings of the Seventh IEEE International Conference on Computer Vision*, pages 1131–1136, 1999.

[48] K Delac, M Grgic, and S Grgic. Independent comparative study of PCA, ICA, and LDA on the FERET data set. *International Journal of Imaging Systems and Technology*, 15(5):252–260, 2006.

[49] A M Martinez and A C Kak. PCA versus LDA. In *EEE Transactions on Pattern Analysis and Machine Intelligence*, volume 23, pages 228–233, February 2001.

[50] L Xiaoguang, W Yunhong, and A K Jain. Combining classifiers for face recognition. In *Proceedings. 2003 International Conference on Multimedia and Expo*, volume 3, pages 13–16, July 2003.

[51] A Hyvrinen and E Oja. Independent component analysis: algorithms and applications. *Neural Network*, 13(4-5):411–430, 2000.

[52] Y Q Zhang, M H Loew, and R L Pickholtz. On modeling the distribution of chest X-ray images and their stochastic properties. In *10th International Conference on Pattern Recognition*, volume 2, pages 218–223, June 1990.

[53] C Liu and H Wechsler. Enhanced Fisher linear discriminant models for face recognition. In *Proceedings. Fourteenth International Conference on Pattern Recognition*, volume 2, pages 1368–1372, August 1998.

[54] R Lotlikar and R Kothari. Face recognition using curvilinear component analysis. In *The 1998 IEEE International Joint Conference on Neural Networks*, volume 3, pages 1778–1783, 1998.

[55] P Navarrete and J Ruiz del Solar. Eigenspace-based recognition of faces: comparisons and a new approach. In *Proceedings 11th International Conference on Image Analysis and Processing*, pages 42–47, 2001.

[56] J R Beveridge, K She, B A Draper, and G H Givens. A nonparametric statistical comparison of principal component and linear discriminant subspaces for face recognition. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 535–542, 2001.

[57] Z Wanfeng, Y Yingchun, and W Zhaohui. Exploiting PCA classifiers to speaker recognition. In *Proceedings of the International Joint Conference on Neural Networks, 2003*, volume 1, pages 820–23, July 2003.

[58] Z Wanfeng, Y Yingchun, and W Zhaohui. A PCA classifier and its application in vehicle detection. In *Proceedings of the International Joint Conference on Neural Networks, 2001*, volume 1, pages 600–604, 2001.

[59] National Electrical Manufacturers Association, 1300 N. 17th Street, Rosslyn, Virginia 22209 USA. *Digital Imaging and Communications in Medicine (DICOM): Version 3.0*, 2001.

[60] W Dillon and M Goldstein. *Multivariate Analysis: Methods and Applications.* John Wiley & Sons, 1984.

[61] W J Krzanowski. *Principles of multivariate analysis.* Oxford science publications, 1988.

[62] G Beltrami. Sulle funzioni bilineari. *Giornale di Mathematiche di Battaglini*, 11:98–106, 1873. Full text unavailable.

[63] M C Jordan. Mémoire sur les formes bilináires. *J. of Math. Pures Appl.*, 19:35–54, 1874. Full text unavailable.

[64] K Pearson. On lines and planes of closest fit to systems of points in space. *Phil. Mag.*, 6(2):559–572, 1901. Full text unavailable.

[65] H Hotelling. Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.*, 24:417–441, 498–520, 1933. Full text unavailable.

[66] Wikipedia. Principal components analysis. http://en.wikipedia.org/wiki/Principal_components_analysis, visited on 1 December 2005.

[67] R B Catell. The scree test for the number of factors. *Multivariate Behavioural Research*, 1:245–76, 1966.

[68] Wikipedia. Mahalanobis distance. http://en.wikipedia.org/wiki/Mahalanobis_distance, visited on 23 July 2006.

[69] R De Maesschalck, D Jouan-Rimbaud, and D L Massart. The Mahalanobis distance. *Chemometrics and Intelligent Laboratory Systems*, 50(1):1–18, January 2000.

[70] D Williams, W H Woodall, J B Birch, and J H Sullivan. On the distribution of the Hotelling's T^2 statistic based on the successive differences

covariance matrix estimator. Technical report, Virginia Polytechnic Institute and State University, 2004.

[71] S Chow. Singular value decomposition and its applications. Master's thesis, University of Canterbury, 1980.

[72] Todd Will. Introduction to the singular value decomposition.
http://www.uwlax.edu/faculty/will/svd/,
Hosted by University of Wisconsin, visited on 15 December 2005.

[73] M Tian, S Luo, and L Liao. An investigation into using singular value decomposition as a method of image compression. In *Proceedings of 2005 International Conference on Machine Learning and Cybernetics*, volume 8, pages 5200–5204, August 2005.

[74] K Konstantinides, B Natarajan, and G S Yovanof. Noise estimation and filtering using block-based singular value decomposition. In *IEEE Transactions on Image Processing*, volume 6, pages 479–483, 1997.

[75] J Gan and Y Zhang. A study of singular value decomposition of face image matrix. In *Proceedings of the 2003 International Conference on Neural Networks and Signal Processing*, volume 1, pages 197–199, December 2003.

[76] D Chandra. Digital image watermarking using singular value decomposition. In *The 2002 45th Midwest Symposium on Circuits and Systems*, volume 3, pages 254–267, August 2002.

[77] H Andrews and C Patterson. Singular value decomposition (SVD) image coding. In *IEEE Transactions on Communications*, volume 24, pages 425–432, April 1976.

[78] H Andrews and C Patterson. Singular value decompositions and digital image processing. In *IEEE Transactions on Acoustics, Speech, and Signal Processing*, volume 24, pages 26–53, February 1976.

[79] The Mathworks Inc. Image processing toolbox 5.1, user's guide. http://www.mathworks.com/access/helpdesk/help/toolbox/images/images.html, visited on 23 January 2006.

[80] Desmond J Higham and Nicholas J Higham. *MATLAB Guide.* Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2000.

[81] M E Wall, A Rechtsteiner, and L M Rocha. *A Practical Approach to Microarray Data Analysis*, chapter 5, pages 91–109. Kluwer: Norwell, MA, 2003.

[82] J J Gerbrands. On the relationships between SVD, KLT and PCA. *Pattern Recognition*, 14:375–381, 3 1981.

[83] A A Maciejewski and R G Roberts. An example of principal component analysis applied to correlated images. In *Proceedings of the 33rd Southeastern Symposium on System Theory*, pages 269–273, March 2001.

[84] E Anderson, Z Bai, C Bischof, L S Blackford, J Demmel, Jack J Dongarra, J Du Croz, S Hammarling, A Greenbaum, A McKenney, and D Sorensen. *LAPACK Users' guide (third ed.)*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1999.

[85] R Weissleder, J Wittenberg, and M Harisingham. *Primer of Diagnostic Imaging.* Mosby, 3rd edition, 2003.

[86] J A Hanley and B J McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143:29–36, 1982.

[87] A. Butler, P. Bones, and M. Hurrell. Eigenimage interpretation of frontal chest radiographs. In *Proceedings RANZCR New Zealand Branch Annual Scientific Meeting*, 2003.

[88] A. Butler, P. Bones, and M. Hurrell. Classification of chest radiographs using eigenimage processing. In *Proceedings RANZCR New Zealand Branch Annual Scientific Meeting*, July 2004.

[89] P. Bones and A. Butler. Eigen-analysis for classifying chest X-ray images. In *Proceedings of the Society for Photo-optical Instrumentation Engineers Vol. 5562 Image Reconstruction from Incomplete Data III*, volume 5562, pages 88–96, 2004.

[90] A. Butler, P. Bones, and M. Hurrell. Eigenimage processing of high resolution chest radiographs. In *Proceedings of the Radiological Society of North America, 90th Scientific Assembly and Annual Meeting*, 2004.

[91] P M Baggenstoss. Class-specific classifier: avoiding the curse of dimensionality. In *IEEE Aerospace and Electronic Systems Magazine*, volume 19, pages 37–52, January 2004.

[92] J P W Pluim and J M Fitzpatrick. Image registration. *IEEE Transactions on Medical Imaging.*, 22(11):1341–1343, November 2003.

[93] J. Maintz and M. Viergever. A survey of medical image registration. *Medical Image Analysis*, 2(1):1–36, 1998.

[94] W Rui and L Minglu. An overview of medical image registration. In *Fifth International Conference on Computational Intelligence and Multimedia Applications.*, pages 385–390, September 2003.

[95] L Ramirez, N G Durdle, and V J Raso. Medical image registration in computational intelligence framework: a review. In *Canadian Conference on Electrical and Computer Engineering*, volume 2, pages 1021–2024, May 2003.

[96] S Periaswamy. *General-purpose medical image registration.* PhD thesis, Department of Computer Science, Dartmouth College, 2003.

[97] B F Hutton and M Braun. Software for image registration: algorithms, accuracy, efficacy. *Seminars in Nuclear Medicine*, 33(3):180–92, 2003.

[98] D L Hill, P G Batchelor, M Holden, and D J Hawkes. Medical image registration. *Physics in Medicine and Biology*, 46(3):R1–45, March 201.

[99] P. Bones, A. Butler, and M. Hurrell. Enhancement of chest radiographs using eigenimage processing. In *Proceedings of the Society for Photo-optical Instrumentation Engineers Vol. 6316 Image Reconstruction from Incomplete Data III*, volume 6316, 63160C, page 12, August 2006.

[100] P A van den Elsen, E J D Pol, and M A Viergever. Medical image matching– a review with classification. *Engineering in Medicine and Biology Magazine, IEEE*, 12(1):26–29, March 1993.

[101] T M Lehmann. A two-stage algorithm for model-based registration of medical images. In *Fourteenth International Conference on Pattern Recognition*, volume 1, pages 344–351, 1998.

[102] Zhao Hui and S Lo. Directional edge registration for temporal chest image subtraction. In *Applied Imagery Pattern Recognition Workshop, AIPR 2001 30th*, pages 32–36, 2001.

[103] M Corvi and G Nicchiotti. Multiresolution image registration. In *International Conference on Image Processing*, volume 3, pages 224–227, 1995.

[104] P Chalermwat and T El-Ghazawi. Multi-resolution image registration using genetics. In *1999 International Conference on Image Processing*, volume 2, pages 452–456, 1999.

[105] J Zhen, J Yifeng, and Z Jihong. Automatic registration algorithm based on wavelet transform. In *5th International Conference on Signal Processing Proceedings*, pages 979–982, 2000.

[106] A Aldroubi and M Unser, editors. *Wavelets in Medicine and Biology*. CRC Press, 1996.

[107] R C Gonzalez, R E Woods, and S L Eddins. *Digital Image Processing using MATLAB*. Upper Saddle River, NJ : Pearson/Prentice Hall, 2004.

[108] S Baker and I Matthews. Equivalence and efficiency of image alignment algorithms. In *Proceedings of the 2001 IEEE Computer Society Conference*

on Computer Vision and Pattern Recognition, volume 1, pages 1090–97, 2001.

[109] S Baker and I Matthews. Lucas-kanade 20 years on: A unifying framework. *International Journal of Computer Vision*, 56(3):221–255, March 2004.

[110] S Baker and I Matthews. Lucas-kanade 20 years on: A unifying framework: Part 1. Technical report, Robotics Institute, Carnegie Mellon University., July 2002.

[111] S Baker, I Matthews, I Matthews, and T Ishikawa. Lucas-kanade 20 years on: A unifying framework: Part 2. Technical report, Robotics Institute, Carnegie Mellon University., February 2003.

[112] S Baker, R Gross, and I Matthews. Lucas-kanade 20 years on: A unifying framework: Part 3. Technical report, Robotics Institute, Carnegie Mellon University., November 2003.

[113] S Baker, R Gross, and I Matthews. Lucas-kanade 20 years on: A unifying framework: Part 4. Technical report, Robotics Institute, Carnegie Mellon University., February 2004.

[114] S Baker, R Patil, K M Cheung, and I Matthews. Lucas-kanade 20 years on: Part 5. Technical report, Robotics Institute, Carnegie Mellon University., November 2004.

[115] S Baker, R Gross, and I Matthews. Image processing toolbox 5.1, user's guide. http://www.ri.cmu.edu/projects/project_515.html, visited on 18 January 2006.

[116] R P Woods, S R Cherry, and J C Mazziotta. Rapid automated algorithm for aligning and reslicing PET images. *Journal of Computer Assisted Tomograph*, 16:620–633, 1992.

[117] R P Woods. Automated image registration. http://bishopw.loni.ucla.edu/AIR5/, visited on 16 December 2005.

[118] Darren Weber. Bioelectromagnetism Matlab Toolbox. http://eeg.sourceforge.net/, visited on 16 December 2005.

[119] S. Periaswamy, J B Weaver, D M Healy Jr., D Rockmore, P J Kostelec, and H Farid. Differential affine motion estimation for medical image registration. In *SPIE's 45th Annual Meeting*, San Diego, CA, 2000.

[120] S Periaswamy and H Farid. Elastic registration in the presence of intensity variations. *IEEE Transactions on Medical Imaging*, 22(7):865–874, July 2003.

[121] S Periaswamy and H Farid. Image registration. http://www.cs.dartmouth.edu/f̃arid/research/registration.html, visited on 1 February 2006.

[122] S Chardon, B Vozel, and K Chehdi. A comparative study between parametric blur estimation methods. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 6, pages 3233–36, 1999.

[123] S Chitale and W T Padgett. Blur identification and correction for a given imaging system. In *Proceedings of IEEE Southeastcon '99*, pages 268–73, 1999.

[124] A Jalobeanu, L Blanc-Feraud, and J Zerubia. Estimation of blur and noise parameters in remote sensing. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 4, pages 3580–3, 2002.

[125] S K Kim and J K Paik. Out-of-focus blur estimation and restoration for digital auto-focusing system. *Electronics Letters*, 34(12):1217–19, 1998.

[126] P Marziliano, F Dufaux, S Winkler, and T Ebrahimi. A no-reference perceptual blur metric. In *Proceedings of 2002 International Conference on Image Processing*, volume 3, pages 57–60, 2002.

[127] P Marziliano, F Dufaux, S Winkler, and T Ebrahimi. Perceptual blur and ringing metrics: Application to jpeg2000. In *Signal Processing: Image Communication*, volume 19, pages 163–172, February 2004.

[128] L Guttman. Some necessary conditions for common-factor analysis. *Psychometrika*, 19(2):159–161, June 1954.

[129] H Kaiser. The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23:187–200, 1958.

[130] I Jolliffe. Discarding variables in a principal component analysis. I: Artificial data. *Applied Statistics*, 21(2):160–173, 1972.

[131] I Jolliffe. Discarding variables in a principal component analysis. II: Real data. *Applied Statistics*, 22(1):21–31, 1973.

[132] C M Waternaux. Asymptotic distribution of the sample roots for a non-normal population. *Biometrika*, 63(3):639–645, 1976.

[133] A W Davis. Asymptotic theory for principal components analysis: non-normal case. *Australasian Journal of Statistics*, 19:206–221, 1977.

[134] W J Conover. *Practical nonparametric statistics*. Wiley, 3rd edition, 1999.

[135] Zhujie and Y L Yu. Face recognition with eigenfaces. In *Proceedings of the IEEE International Conference on Industrial Technology, 1994*, pages 434–438, 1994.

[136] A Pentland, B Moghaddam, and T Starner. View-based and modular eigenspaces for face recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Seattle, WA, June 1994.

[137] T E Keats, W Mark, and M D Anderson, editors. *Atlas of Normal Roentgen Variants That May Simulate Disease*. Mosby, 7th edition, 2001.

[138] R G Grainger, D J Allison, A Adam, A K Dixon, H Carty, and Alan Sprigg, editors. *Grainger & Allison's Diagnostic Radiology: A Textbook of Medical Imaging*. Churchill Livingstone, 4th edition, 2001.

[139] N.L Herbst, B.M. Muller. Building a representative training set based on eigenimages. In *Fourteenth International Conference on Pattern Recognition*, volume 2, pages 1846–1848, Aug 1998.

[140] M Artac, M Jogan, and A Leonardis. Incremental PCA for on-line visual learning and recognition. In *16th International Conference on Pattern Recognition*, volume 3, pages 781–784, 2002.

[141] H Zhao, P C Yuen, J T Kwok, and J Yang. Incremental PCA based face recognition. In *Control, Automation, Robotics and Vision Conference*.

[142] Y Li, L Q Xu, J Morphett, and R Jacobs. An integrated algorithm of incremental and robust PCA. In *International Conference on Image Processing*, volume 1, pages 245–8, September 2003.

[143] K Nishino, S K Nayar, and T Jebara. Clustered blockwise PCA for representing visual data. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 27, pages 1675–1679, October 2005.

[144] P Hall, D Marshall, and R Martin. Incremental eigenanalysis for classification. In John Carter and M Nixon, editors, *Proceedings of the Ninth British Machine Vision Conference*, pages 286–295, 1998.

[145] P Hall, D Marshall, and R Martin. Merging and splitting eigenspace models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(9):1042–1049, September 2000.

[146] P Hall, D Marshall, and R Martin. Adding and subtracting eigenspaces with eigenvalue decomposition and singular value decomposition. *Image and Vision Computing*, 20(13-14):1009–1016, December 2002.

[147] A Pentland, T Starner, N Etcoff, A Masoiu, O Oliyide, and M Turk. Experiments with eigenfaces. In *International Joint Conference on Artificial Intelligence*, August 1993.

[148] A. Butler, P. Bones, and M. Hurrell. A method to enhance frontal chest radiographs in the eigen-domain. In *Proceedings of Image and Vision Computing New Zealand, 2005*, pages 363–368, 2005.

[149] H Soltanian-Zadeh, J P Windham, and J M Jenkins. Error propagation in eigenimage filtering. In *IEEE Transactions on Medical Imaging*, volume 9, pages 405–420, December 1990.

[150] H Soltanian-Zadeh, A E Yagle, J P Windham, and D O Hearshen. Optimization of MRI protocols and pulse sequence parameters for eigenimage filtering. In *Nuclear Science Symposium and Medical Imaging Conference*, volume 2, pages 1325–1327, October 1992.

[151] Liu Rujie; Yuan Baozong. Eigenspace-based human face detection. In *5th International Conference on Signal Processing*, pages 1305–1308, 2000.

[152] N Vaswani and R Chellappa. Principal components null space analysis for image and video classification. In *IEEE Transactions on Image Processing*, volume 15, pages 1816–1830, July 2006.

[153] J E Jackson and R H Morris. An application of multivariate quality control to photographic processing. *Journal of the American Statistical Association*, 52(278):186–199, June 1957.

[154] Michael Sandborg, Anders Tingberg, Gustaf Ullman, David Dance, and Gudrun Alm Carlsson. Comparison of clinical and physical measures of image quality in chest and pelvis computed radiography at different tube voltages. *Medical Physics*, 33(11):4169–4175, November 2006.

[155] C Lee, H Tschai, Y Kuo, W Ko, Y Cheng, and C Wang. PACS: construction and application to medical image enhancement. In *Third International*

Conference on Knowledge-Based Intelligent Information Engineering Systems, pages 246–249, December 1999.

Appendix A

Publications

A.1 Citable papers

1. A Butler, P Bones, and M Hurrell. A prototype system for enhancement of frontal chest radiographs using eigenimage processing. *Australasian Radiology*, Accepted, pending publication.
   (This is the journal of the Royal Australian and New Zealand College of Radiologists. It is Australasia's largest and most widely read radiology journal.)

2. P Bones and A Butler. Enhancement of chest radiographs using eigenimage processing. In *Proceedings of the Society for Photo-optical Instrumentation Engineers Vol. 6316 Image Reconstruction from Incomplete Data III*, 2006. ref. 63160C, 12 pp.
   (This paper was presented by Associate Professor Bones on behalf of the team, at the SPIE conference on image reconstruction in San Diego. This is an important international conference for image processing.)

3. A Butler, P Bones, and M Hurrell. A method to enhance frontal chest radiographs in the eigen-domain. *Proceedings of Image and Vision Computing New Zealand, 2005*, pp 363-368.
   (This is New Zealand's largest Computer Vision conference and covers many aspects of automated pattern recognition.)

4. P Bones and A Butler. Eigen-analysis for classifying chest X-ray images. In *Proceedings of the Society for Photo-optical Instrumentation Engineers Vol. 5562 Image Reconstruction from Incomplete Data III*, 2004. pp 88-96.

(This paper was presented by Associate Professor Bones on behalf of the team, at the SPIE conference on image reconstruction.)

A.2   Abstracts in scientific proceedings

1. A Butler, P Bones, and M Hurrell. Enhancement of frontal chest radiographs using eigenimage processing. In *Australian Radiology, Vol. 50, Sup. 1*, 2006. pp A6.
   (This is the largest and most prestigious radiological conference in Australasia. This abstract and presentation won the "Philips Medical Systems Radiology Paper Prize".)

2. A Butler, P Bones, and M Hurrell. Eigenimage Processing of High Resolution Chest Radiographs. In *Proceedings of the Radiological Society of North America, 90th Scientific Assembly and Annual Meeting*, Dec. 2004.
   (This is the largest and most prestigious radiological conference in the world, attracting more than 60,000 delegates annually.)

3. A Butler, P Bones, and M Hurrell. Classification of Chest Radiographs using Eigenimage Processing. In *Proceedings RANZCR New Zealand Branch Annual Scientific Meeting*, Jul. 2004.
   (This is the premier New Zealand scientific conference for radiology.)

4. A Butler, P Bones, and M Hurrell. Eigenimage Interpretation of Frontal Chest Radiographs. In *Proceedings RANZCR New Zealand Branch Annual Scientific Meeting*, Jul. 2003.
   (This is the premier New Zealand scientific conference for radiology.)

A.3   Publicity and popular press

1. "Canterbury Medical Research Foundation Open Day", Aug. 28, 2005
   (This is Canterbury's largest open day for Medical Researchers to present their work to the public. I represented the radiology research group talking to members of the public about current radiology research.)

2. "Radiologist's university research to benefit his daily work"
   "The Chronicle," Oct. 20, 2005. (Vol 40, num 17.)
   (This is the University of Canterbury Staff newsletter. The article detailed the "GE-Medical Systems Prize".)

Appendix B

Awards

1. "Health Research Council, Clinical Research Training Fellowship", 2006.
   (This fellowship provides a stipend for a medical specialist for up to three years full time study to complete a Ph.D. It was awarded for the final year of my thesis.)

2. "R. H. T. Bates Postgraduate Scholarship", 2006.
   (This scholarship is awarded by the Royal Society of New Zealand in memory of Professor Bates. It is awarded to a Ph.D. student whose research aims to apply information theory or image processing to studies in medicine, the physical sciences, astronomy or engineering. No financial benefit was available from this award due to conflict with the HRC award.)

3. "Philips Medical Systems Radiology Paper Prize", 2006.
   (This prize is for the best presentation by a junior radiologist in the radiology section of the annual scientific meeting of the Royal Australian and New Zealand College of Radiologists. A junior radiologist is either a student member or a fellow who has held their specialist qualifications for not more than four years.)

4. "Entré 10k Challenge", 2006.
   (In this competition teams prepare a short venture summary on their enterprising idea. Julian Maclaren and myself prepared a business plan which focussed on commercialising eigenimage processing. There are ten equal winners for this competition.)

5. "GE Medical Systems Prize", 2005.
   (This is the top research prize of the Royal Australian and New Zealand

College of Radiologists. It is awarded to the best research proposal by a student member or fellow of less than ten years.)

6. "Obex Medical Education Award: South Island Radiology Registrar Award", 2004.
   (This prize is for the best presentation of original research by a South Island Radiology Registrar.)

7. "Obex Medical Education Award: South Island Radiology Registrar Award", 2003.
   (This prize is for the best presentation of original research by a South Island Radiology Registrar.)