

AUTOMATION IN MAMMOGRAPHY : COMPUTER VISION AND HUMAN PERCEPTION

Sue Astley*, Ian Hutt*, Stephen Adamson†, Peter Miller*,
Peter Rose*, Caroline Boggis‡, Chris Taylor*, Tim Valentine†,
Jack Davies* and Janette Armstrong*

*Departments of *Medical Biophysics and †Psychology, University of Manchester, Oxford Road, Manchester M13 9PT, UK, ‡The North West Regional Training Centre for Breast Screening, Withington Hospital, Nell Lane, Manchester, and *The Wolfson Breast Pathology Unit, Southmead Hospital, Bristol.*

ABSTRACT

Mammographic screening programmes generate large numbers of highly variable, complex images, most of which are unequivocally normal. When present, abnormalities may be small or subtle. Two processes critical to the success of screening programmes are the perception of potential abnormalities and the subsequent analysis of each detected lesion to determine its clinical significance. The consequences of errors are costly, and in many screening centres, films are read by two radiologists in an attempt to reduce errors. The prime objective of our research is to improve the accuracy of the detection and analysis of breast lesions by providing radiologists with computer-aided digital image analysis tools. In this paper we focus on the detection and analysis of mammographic microcalcifications.

We describe a philosophy of research aimed at generating useful computer-based aids for radiologists. Firstly, it is necessary to accurately identify specific tasks which are difficult for the human observer. Having correctly identified a problem, appropriate computer vision methods must be developed and their performance evaluated. It is then important to determine effective ways of using such methods to aid radiologists, and it is essential to prove that the effect on radiologists' performance is entirely beneficial.

We present results of experiments to determine factors affecting radiologists' perception of microcalcifications, and to investigate the effects of attention-cueing on detection performance. Our results show that radiologists' performance can be significantly improved with the use of prompts generated from automatically-detected microcalcification clusters.

We describe a new method for the delineation of mammographic abnormalities based on the analysis of multiple high quality X-ray projections of excised lesions. Biopsy specimens are secured inside a rigid tetrahedron, the edges of which provide a reference frame to which the locations of features can be related. A three-dimensional representation of an abnormality can be formed and rotated to resemble its appearance in the original mammogram.

1. INTRODUCTION

There is now unequivocal evidence that the mortality from breast cancer can be reduced by mammographic screening¹⁸; consequently, a number of organised

screening programmes based on X-ray examination of the breasts (mammography) have been instituted. For example in Britain, all women between the ages of 50 and 65 are invited to join a national programme which currently involves single-view medio-lateral oblique mammography once every three years. At present, this programme generates over 1.5 million mammograms per year, most of which are interpreted at a small number of specialised centres distributed around the country²².

In the UK, screening mammograms are carefully searched for any signs of abnormality by experienced breast radiologists, who are expected to maintain their level of expertise by interpreting at least 6,000 mammograms every year³⁸. Despite a strong emphasis on training, practice and experience, radiologists still find mammographic interpretation demanding; both normal and abnormal appearances are highly variable and often complex, and clinically important lesions may be small or subtle. There is a significant level of intra- and inter-observer variability¹⁹, and in many centres mammograms are read independently by two radiologists in an attempt to reduce errors.

Radiologists look for three major classes of mammographic abnormality; discrete abnormalities, diffuse changes, and alterations between successive films of the same breast¹⁸. Discrete abnormalities include clusters of microcalcifications, local opacities, and localised distortion of the normal structure of the breast. Diffuse changes may manifest themselves as asymmetry between right and left breast images, or as widespread calcification.

A process central to mammographic interpretation is the perception of potential abnormalities. Individual particles of microcalcification as small as 0.1mm can be detected mammographically, although they are only considered to be clinically significant if they appear in clusters of three or more particles⁴². Invasive tumours must be detected before they exceed about 1cm in diameter if screening is to be successful¹⁸. The problem of perception is compounded by the fact that abnormalities are relatively infrequent; in the British programme, less than 1.5% of women screened have mammographic abnormalities considered sufficiently suspicious to warrant biopsy¹⁷.

Another critical aspect of mammographic interpretation is the analysis of each potential abnormality to determine its clinical significance. Analysis is more likely to be a problem for relatively inexperienced interpreters, or those working in smaller practices, who may not have come across sufficient examples of different lesion types to be able to reliably distinguish them from one another. In the UK approximately two benign lesions are biopsied for every one malignant lesion¹⁷; clearly, it would be advantageous to reduce this ratio, provided no genuine malignancies would then be missed.

Computer-based image processing and analysis tools are now commonplace in radiology departments; for example, windowing facilities are routinely used to interactively change the appearance of CT images, and echocardiographers regularly make measurements of ventricular volumes by indicating key points on the images with a cursor. Benefits of computer-based techniques such as these include the ability to make accurate, objective, reproducible measurements.

We are investigating the application of computer vision techniques to mammography, with the aim of improving the accuracy of interpretation. Although we would ideally like to construct a system in which all forms of abnormality are automatically detected and classified, the current state of the art in mammographic image analysis renders this a long-term goal. We are therefore developing computer-based aids which will be useful in their own right, but will also eventually contribute to a completely automated system. We have considered a number of possible modes of assistance: enhancement, in which abnormalities are made more conspicuous by enhancing diagnostically significant features and suppressing insignificant features; prompting, in which image features related to abnormalities are detected automatically and then used to draw the observer's attention to suspicious locations; and analysis, in which properties of potentially abnormal regions are extracted to determine whether or not the abnormality is genuine, and to allow classification, where appropriate.

In this paper, we selectively examine the state of the art both in the automated detection and analysis of mammographic abnormalities, and in the application of these methods in a clinical environment. We then describe our own strategy for developing practical, realistic, computer-based aids for radiologists engaged in mammographic interpretation. This is viewed in the context of a functional description of a system which might realistically be of benefit to radiologists in the short term. We identify the gaps in our knowledge which prevent us from constructing such a system immediately, and those experiments essential to a thorough investigation of feasibility. We describe our own attempts at answering some of these questions, focusing on the detection and analysis of microcalcifications.

2. BACKGROUND

Much of the published research in automated mammographic interpretation has focussed on the detection of microcalcifications. Microcalcifications are a natural place to start research in this field; they are one of the earliest signs of breast cancer and, although small, they are easy to describe and relatively easy to distinguish from normal breast structures.

Several apparently successful techniques for detecting microcalcifications have been described in the literature^{9,12,25} etc., though none have yet been tested thoroughly. Most methods apply a sequence of progressively more sophisticated operations to an image to generate and refine a set of candidate abnormalities. The properties of these candidates are then measured, and pre-determined thresholds decide which candidates merit further investigation. A major problem with this type of approach is that any errors in the earlier stages of the detection process are propagated through to later stages. Karrsemeijer has described a probabilistic approach which overcomes this difficulty by generating separate parametric images to represent local shape and contrast, and by incorporating continuity and clustering constraints in an iterative scheme²⁵. His method is theoretically attractive, but computationally expensive. Our own approach to detecting microcalcifications, which we discuss later in this paper, is also based on Bayesian statistics². Bourrely and Muller have demonstrated the use of a

neural network to discriminate mammographic microcalcifications from normal background structures⁴. Their results are promising, with relatively low false negative rates, although the high degree of background variability results in a large proportion of image regions remaining unclassified by the network.

The automated classification of microcalcifications has also been investigated with moderate success by researchers aiming to identify characteristic properties of benign and malignant individual calcifications and calcification clusters e.g. ^{21,30}. There is general agreement that one of the most salient properties in discriminating benign from malignant clusters is the number of particles in close proximity. Lanyi³⁰ has performed extensive studies of the shapes of clusters of microcalcifications. His results suggest that the discrimination between benign and malignant lesions can be greatly improved by evaluating both the shapes of individual microcalcifications and the configurations of clusters. A significant drawback of his approach is that the apparent shape of a cluster will change with X-ray projection.

Automated tumour detection has been less widely researched and produced less impressive results, partly because of the wide variability in tumour appearance, and the visual similarity of some tumours to normal structures in dense and fatty-glandular breasts. Approaches range from a comparison of left and right breast images^{20,31,35} to indirect detection by searching for radiating patterns of lines characteristic of spiculation and architectural distortion^{3,26}. To date, few researchers have attempted a thorough assessment of performance.

Giger²⁰, Kimme²⁷, Hoyer²³ and Semmlow³⁹ have used asymmetry between left and right breasts as an initial cue. These researchers all failed to deal satisfactorily with the problem of differences in size and shape of the two breasts; Hoyer and Kimme applied arbitrary partitioning to obtain approximate correspondence, whilst Giger matched the breast boundaries as far as possible and then performed a direct subtraction, ignoring any potential abnormalities close to the skinline. Our strategy for the detection of asymmetry is based on the comparison of regions of similar composition, detected by texture analysis^{35,36}. Alternative approaches to lesion detection have been proposed by Lai²⁹, Brzakovic⁵ and Kegelmeyer²⁶. Lai's computationally expensive method involved template matching at a range of resolutions, and Brzakovic's method was based on a multi-resolution analysis of image texture; neither produced clinically acceptable error rates. Kegelmeyer has investigated the detection of spiculated lesions by the analysis of locally oriented edges and by texture analysis; his initial results, using a limited test set of five images, are encouraging. We have also performed preliminary experiments on the detection of spiculated lesions, using a technique based on the Hough transform to detect characteristic properties of radiating linear structures³. Our method performs well for relatively uniform parenchymal patterns, although an extensive evaluation has not yet been performed.

Radiographic enhancement has been described by a number of authors^{8,13,11,16} etc., although few have attempted to assess the clinical impact of such techniques. Enhancement falls into two broad categories; general improvement of image appearance^{11,16}, and enhancement of specific features associated with disease. For example, Chan⁸ employed an unsharp-mask filter to selectively enhance microcalci-

cation-like features and found that this led to improvements in radiologists' detection performance when compared to their performance viewing the unprocessed digital image, although the best results were still obtained using the original, undigitised image. Dhawan¹³ investigated a range of contrast enhancement procedures based on optimal adaptive neighbourhood processing. Unfortunately, his results were not presented to any radiologists to determine whether the procedures actually led to improvements in the detectability of clinically significant structures. We are currently performing a series of experiments designed to test the effect of enhancement of lines, edges and small peaks on radiologists' perception of subtle mammographic abnormalities. Our enhancement operator is Dixon's line detector¹⁴; it is applied to a gaussian pyramidal representation of the mammogram⁶. These experiments use 156 normal and abnormal mammograms, each presented in the original form, as an unenhanced digital image printed on film, and in an enhanced form also printed on film. Results of the full experiments are not yet available, but an initial feasibility study has indicated that use of Dixon's operator might partially overcome the degradation in performance noted when radiologists are required to make diagnoses from digital images.

Another application of computer vision methods in mammography is the analysis of parenchymal patterns. Based on the premise that breast patterns may be related to the natural risk of a woman developing breast cancer, a number of researchers have attempted, with varying degrees of success, to automatically classify patterns into those designated by Wolfe⁴⁴ as being clinically significant. The identification and analysis of glandular patterns is also useful from a dosimetry standpoint, since the risk incurred by screening depends partly on the glandular composition of each woman's breasts. Although Magnin's analysis of textural features did not appear to yield any effective way of discriminating between parenchymal pattern types³⁴, Caldwell's method, based on fractal analysis, did demonstrate a classification agreement between the system and a group of radiologists that was only slightly lower than the agreement between the radiologists⁷. Similarly, Shadagopan attempted to quantify duct patterns and obtained a good correlation between the computer's calculations and the measurements made by a human observer, although the data set used in this study was limited⁴¹.

It has been reported that significant levels of intra- and inter-observer variability exist in mammography¹⁹. False positives can be determined by studying screening centre audit data such as the ratio of benign to malignant lesions biopsied. False negatives (cancers missed due to observer error) can be identified by studying the so called interval cancers found in the more mature screening programmes. In a study of interval cancers over during twelve years of the Nijmegen programme³⁷ it was found that 26% of these cancers were actually missed at the previous screening examination because of technical or observer error. This is about half of the number of genuine interval cancers, that is, cancers arising within the two year screening interval. It was found that, of all cancers detected at screens other than the initial screen, approximately 20% could, in retrospect, be detected in a previous mammogram.

Kundel and Nodine have studied the search behaviour of radiologists scanning for small lung abnormalities in chest X-rays²⁸, a task analogous to the detection of small abnormalities in mammograms. They observed that the scanning patterns employed by these radiologists were neither systematic nor complete, and that the pattern of fixations could be influenced by the provision of specific clinical information regarding the patient prior to the presentation of the film. Kundel and Nodine also investigated the types of errors made by radiologists during the film reading process; their results suggest that around half of the errors were due to an insufficient level of attention being directed towards the location of the abnormality. A small but important literature is devoted to assessing the effect of computer-based tools on the performance of human interpreters. Experimental evidence suggests that prompting can improve human detection performance in highly controlled visual search tasks, by directing observers' attention towards targets⁴³; this has been investigated to a limited extent in relation to the perception of mammographic abnormalities¹⁰. In this paper we describe results of experiments which provide independent confirmation that prompting can indeed be beneficial in improving radiologists' detection of subtle mammographic abnormalities.

3. A FRAMEWORK FOR ASSISTANCE

To date, research into computer-based mammographic interpretation has been performed in a largely uncoordinated manner. Good progress has been made in some areas, whilst potentially successful methods have still not been found for the more difficult, subjective manifestations of abnormality. Few methods have been rigorously tested, and proven to be of clinically acceptable standard. Here we describe a framework in which both new and existing computer-based mammographic interpretation tools could be placed; we then modify this description to a more immediately realistic level.

Ideally, our system would perform the following functions:

- performance monitoring
- image acquisition from digitised film images and from digital radiography systems
- data compression
- image display and manipulation
- image restoration
- image enhancement
- pre-screening
- prompting
- analysis and classification
- on-line assistance
- reporting
- teaching and research

The digitisation system for acquiring mammograms from film should be flexible and relatively fast, producing high resolution, high quality images. It should support

automatic input of films, possibly reading bar-codes to obtain film identification and patient data. Rapid, loss-free data compression will be essential to enable effective transmission, storage and retrieval of images. Quality control is vital for successful mass screening programmes; our system should monitor factors such as patient positioning, image quality, and interpretation standards, with the aim of identifying and correcting any problems at an early stage. Remedial action to correct defects or distortions due to the imaging and acquisition process can also be taken, to minimise the number of repeat mammograms required.

A variety of hard-copy and soft-copy display mechanisms and image manipulation tools should be provided to allow control over both the appearance of the image and the nature of the display. We assume that the system will be used primarily by radiographers and radiologists, so it must have a simple, adaptable user interface that can be tailored to individual user's requirements. In addition to standard functions such as contrast enhancement, windowing, pan and zoom, more application-specific display tools can be provided. For example, one method of facilitating the comparison of right and left breast images would be to horizontally reverse one of the pair and then register and rapidly toggle between the two images. A similar technique without reversal could assist the assessment of subsequent examinations of the same breast. Similarly, enhancement of image features specifically associated with signs of abnormality might aid the perception process.

Ideally, we would like our system to perform pre-screening, that is, to automatically categorize mammograms into groups such as 'definitely normal', 'definitely abnormal', 'equivocal', and 'technical failure'. This requires reliable detection of all manifestations of mammographic abnormality, since one cannot classify a mammogram as 'definitely normal' without eliminating the possibility that it contains any abnormality. Clearly, this is a much more difficult task than assigning a film to one of the other three categories. However, if our system is able to detect specific types of abnormality, it can be used to 'prompt' radiologists by indicating to them any locations deemed to be suspicious. There are two main requirements for the development of a prompting system; at least one method for detecting suspicious regions, and the ability to present the information to the radiologist in a way which will be helpful.

Once an abnormality has been detected either by human or machine, it must be analysed to determine its clinical significance and, if possible, to make a tentative diagnosis. There are many facets to analysis, including the extraction of quantitative descriptions of radiological properties such as size, shape and density, and comparison with models of known lesion types. Such models incorporate prior knowledge of lesion properties and of the radiological process, and can in part be generated from example lesions. Our system may also provide more specific analytical assistance including: lesion measurements, to enable the accurate assessment of the efficacy of treatment; estimates of the risk associated with screening an individual, based on measurements of the amount of parenchymal tissue; delineation of abnormalities for the planning of further investigation and treatment; and assistance with relating features in multiple X-ray projections of the same breast. A library of example lesions and other abnormal

features can also be made available to refresh radiologists' memory of less common abnormalities.

Such a system will inevitably lead to the establishment of a database not only of images, but of image and lesion features, and of radiologists' annotations. Each new proven example of an abnormality can be added to the database, and incorporated into the appropriate statistical models to improve model reliability in capturing individual parameters and their variability. The database will be invaluable both for research and teaching purposes. The system will itself be a useful training aid for mammographic interpreters given the provision of suitably structured access to the database of images, experienced radiologists' annotations and comments, and pathological information about each case.

A final important stage in the interpretation of mammograms is reporting. There are considerable advantages in using a computer-based system for this purpose; to date, most reports incorporate a highly stylized sketch diagram on which the radiologist can indicate the location of any abnormality. With digital images, it will be possible to produce a diagram based on a reduced version of the original mammogram accurately annotated by the radiologist. Such a diagram will provide detailed, accurate information for those involved in treating any abnormalities detected.

Having described the functionality of an ideal computer-based mammographic interpretation system, we now consider how such a system might be realised in practice, bearing in mind both technical issues and other important factors such as cost-effectiveness and time-scale for development. From our analysis of the state of the art in computer-based mammographic image analysis, we can clearly see that the production of a comprehensive system such as that described above is infeasible at the present time. The most significant difficulty lies with automated pre-screening, since reliable methods for the detection of *all* mammographic signs of abnormality have not yet been developed. Another major problem lies in the engineering of the system; the handling of large numbers of X-ray films, digitisation to sufficiently high quality at reasonable speed and cost, and dealing with copious quantities of digital data.

A more realistic target in the short term would be a system to aid radiologists by providing basic functions such as image restoration, image manipulation and display tools, the automatic detection of a limited range of specific types of abnormality, prompting of these abnormalities, and analysis. Such a system could be implemented incrementally, with further functionality added as it becomes available. Problems still exist in system engineering terms, but we are no longer faced with the daunting task of finding and validating a method for every possible abnormal appearance. In practice, the system could be used to detect abnormalities and generate prompts overnight; in the day-time, radiologists could use it interactively for image enhancement, prompting, analysis and reporting.

4. DETERMINING FEASIBILITY : THE MANCHESTER APPROACH

Although considerable progress has been made by ourselves and others developing computer-based mammographic analysis methods, a number of important issues are outstanding. Firstly, we must question our motivation for developing particular methods. Are we actually focussing our attention on areas where human interpreters need assistance? Secondly, we need to know how well our methods must perform to provide useful assistance. For example, if a detection method is to be used as a basis for prompt generation, we must establish the effects of different types of prompting error on radiologists' performance. Thirdly, we must be aware of possible side-effects. For example, if we prompt observers with the locations of one type of automatically detected abnormality, we must determine any effect these prompts might have on the observers' detection of other signs of abnormality. These issues can be resolved by performing experiments to investigate the behaviour of experienced human interpreters.

In this section, we describe progress which we have made in Manchester towards investigating the feasibility of computer-aided mammographic interpretation. Our general strategy is as follows:

- find out what detection and analysis tasks are problematic for radiologists
- establish a framework for measuring performance
- develop computer-based methods for detection and analysis
- find out how these methods can be used effectively

We now address each of these areas in turn.

4.1 Identify problematic tasks

Two important requirements for effective breast screening are accurate, consistent detection and analysis of mammographic abnormalities, and the efficient use of human experts' time. Both of these areas can be addressed using computer vision methods; we must therefore determine the precise nature of problems encountered by human interpreters, and target our research appropriately.

More specifically, we must identify aspects of mammogram interpretation which are:

- perceptually difficult
- analytically difficult
- tedious
- time-consuming

Signs of abnormality which are difficult to perceive are natural candidates for computer-based enhancement, detection and feature extraction. Both normal and abnormal features can pose analytical problems; they may be subjective or ill-defined, highly variable, complex or rare. It may thus be difficult for radiologists to learn their characteristic patterns and to achieve accurate, consistent assessment. Tasks which are tedious are more likely to give rise to inconsistent performance, whilst those which are relatively easy but time-consuming are expensive in radiologists' time.

In many cases it is easy to suggest plausible candidates for these types of tasks. Microcalcifications are small, and may have poor contrast (i.e. are perceptually difficult). Asymmetry and subtle distortions of breast architecture are analytically difficult, being variable, ill-defined and subjective. The screening process as a whole can be tedious and time-consuming, as the vast majority of films are unequivocally normal. However, these suggestions are based mainly on intuition and on knowledge of the visual task. A more correct approach is to examine 'missed' abnormalities, that is, interval cancers and inconsistencies which come to light in double reading, and to relate radiological diagnoses to pathological data. Experiments with human observers and carefully controlled visual stimuli can be used to confirm or refute hypotheses about which tasks are problematic, and why.

One such hypothesis is that the density and complexity of the glandular pattern will influence the ability of radiologists to detect subtle features; if this is indeed the case, we can target our assistance at those types of image which cause the most problems. We have performed experiments to test this hypothesis in the case of microcalcification clusters, using synthetic abnormalities superimposed on digitised normal mammograms¹. These images, and a selection of normals, were presented briefly to observers who were required to detect and localise any abnormalities.

Our data comprised a set of 15 normal mammograms, classified by an experienced breast radiologist into three categories; fatty-homogeneous, fatty-glandular, and dense-homogeneous. A 6.5cm square region of each image was digitised to an effective pixel size of 0.3mm, with 8 bit grey resolution. The mammograms were magnified by a factor of five to ensure that the natural size relationship between microcalcifications and normal structures was maintained, since the minimum size of the synthetic microcalcifications was determined by the spatial resolution of the viewing device (a high resolution computer screen). In each case, the region digitised was selected to cover the area of the mammogram in which microcalcifications are most likely to arise. No obvious orienting features such as the skin-line or the pectoral muscle were included. The image set was expanded to a total of 60 images by image duplication. Images were randomly allocated to normal and abnormal groups of equal size. One third of all the images were rotated by 90° clockwise and one third by 90° anti-clockwise to produce a greater diversity of appearance.

Artificial clusters of microcalcifications were generated graphically, using the following constraints:

- 3 to 8 particles per cluster
- 1 to 5 pixels per particle
- particle grey level values in top 21.5% of range

Within these constraints, the assignment of particle shapes, sizes and grey levels was random. The number of pixels in each cluster and the locations of individual particles within a bounding box 25 pixels in diameter were also assigned randomly. Each image was partitioned into nine square response regions. Each cluster was placed manually in a realistic location on an 'abnormal' image, avoiding the boundaries between response regions. A training data set was produced in a similar fashion, using different

mammograms to produce fifteen images with which subjects could familiarise themselves with the protocol. Our five observers were all experienced breast radiologists.

Observers were repeatedly presented with the following sequence of images: a response grid with the cursor positioned at the centre to provide a fixation point (1 second); the normal or abnormal mammogram (2, 5 or 8 seconds); the response grid, in which the observer was required to mark the location of any cluster (unlimited). The images were randomised, and divided into three blocks, corresponding to the three mammogram presentation times. Each observer was assigned a block order with either decreasing or increasing presentation time, since true counterbalancing could not be achieved with just five observers.

Observers were instructed to indicate their responses using one of the following criteria; a cluster is definitely present, a cluster is probably present, a cluster is possibly present, or the image contains a suspicious region. Our results showed that the criterion used had no significant effect on performance. We believe that, despite our written request, the observers did not actually adopt different criteria, since there was no powerful incentive for them to do so.

The variation of presentation rate between two and eight seconds also had little effect on performance, although a trend was observed in $P(TP)$, the probability of detecting a genuine cluster. At two seconds the mean over the five observers was 0.69, rising to 0.78 at five seconds and 0.90 at eight seconds. More surprisingly, in these experiments no significant relationship was found between background type and missed clusters (false negatives). Possible reasons for this are discussed later.

Initially, we found that detection performance was related to cluster location. In particular, there was a significant relationship between missed clusters and their locations in the images ($F_{\text{observed}} = 3.30$, $F_{\text{critical}} = 3.13$ at $P < 0.01$, $df = 8$). We identified a number of factors which might have contributed to this effect. In particular, the limited size of our data set, the relationship between its members, and the manual selection of 'realistic' locations rather than entirely random cluster placement, could all be problematic. For example, realistic cluster locations are most likely to reside within the gland, and in a predominantly fatty image with a limited glandular component, the number of such locations will be small. We used only five images of each glandular type; if the glandular regions in these images were biased towards particular squares, rotation to produce additional images would increase the effect of the bias. However, an analysis of false negatives (missed clusters) shows that approximately one third of the total number of missed clusters were missed by three or more of the radiologists, suggesting that these errors might have been caused by physical properties of the stimulus. When these are excluded from the location analysis, no significant relationship between cluster location and missed clusters is found.

In order to investigate the effects of local stimulus properties on detection, true positive locations were defined where all the radiologists correctly located a cluster, false positives where any radiologist mis-classified a normal region, and false

negatives where clusters were missed by more than three observers. The mean grey level standard deviation was measured in 25 pixel square regions corresponding to each of these cases, and to representative normal regions. It was found to have a value of 7.4 in normal background regions, 10.6 in regions where spurious detections were made, 12.8 for missed clusters, and 13.9 for correctly detected clusters. In the false negative cases (missed clusters) it seems likely that genuine clusters were masked by the natural variation of the surrounding structures. A measure of contrast defined by the difference in mean grey levels between calcification and non-calcification pixels, scaled by the standard deviation, was used to investigate any further difference between those clusters which radiologists were able to detect, and those which they failed to detect. There was indeed a significant difference in contrast between the two ($T\text{-observed} = 3.20$, $T\text{-critical} = 2.80$ at $P < 0.01$), with the true positives having higher contrast. In addition, we found that the grey level standard deviations of regions surrounding correctly identified clusters were significantly lower ($T\text{-observed} = 3.35$, $T\text{-critical} = 3.11$ at $P < 0.01$) than those of regions surrounding missed clusters, again implying a masking effect.

In summary, clusters were successfully detected by observers when the surrounding background was relatively uniform, and where the cluster had high contrast. Clusters were missed when the surrounding tissue was non-uniform, and where the cluster had low contrast with its surroundings. This suggests that developing methods to classify images into radiologist-defined glandular pattern categories is unlikely to meet with success, whereas measuring local image properties may be a valid mechanism to guide search for microcalcifications.

4.2 Establish a framework for measuring performance

A major problem for those developing computer-based interpretation tools lies in the difficulty in demonstrating that methods work to an acceptable standard of performance. In this section we describe the various aspects of this problem, and suggest some possible solutions.

One of the first difficulties we encounter is the selection of data for our experiments. If our techniques are to be used for breast screening, we should ideally select data which is representative of that generated by screening programmes. Screening data has two important properties; firstly, the vast majority of films are normal, and secondly, abnormalities are often small or subtle. It is generally necessary to bias the data set to include multiple examples of an abnormality; however, a pitfall of many researchers is that methods tend to be evaluated using only normal films and films showing the type of abnormality the method is tuned to detect. We rarely see what effect a spiculated lesion, for example, might have on the performance of a method for detecting microcalcifications. In addition, any bias of data sets should be declared prior to any psychological experiments, so the results are not influenced by radiologists' prior expectation of the number and type of abnormalities likely to be present.

The size and subtlety of screen-detected lesions is less likely to be problematic, since some relatively advanced lesions can be expected in the first round of screening.

However, data sets should reflect the probable bias to small, subtle abnormalities. The exception to this rule is in the selection of data for preliminary and exploratory experiments, such as those investigating the effects of image enhancement. In this case it is valuable to include both subtle and 'ball-park' examples in the data set. It is also important to ensure that the natural variability both of lesion and background appearances is captured in the data set. We have adopted a matrix approach to selecting examples; the matrix has dimensions representing glandular pattern (predominantly fatty, mixed fatty-glandular, predominantly glandular), abnormality type (no abnormality, calcification, well-defined lesion, ill-defined lesion, spiculated lesion, asymmetry, architectural distortion), and diagnosis type (benign, equivocal, malignant). In this matrix, subtlety is represented in the third dimension, in which lesions of known pathology that are difficult to classify radiologically are placed in the equivocal group.

Another important issue is how to accurately identify and delineate normal and abnormal structures visualised in mammograms. This is essential for both training and test purposes. The most reliable source of evidence about a given region in a mammogram is pathological investigation; this is rarely practical, and while it may be appropriate for characterising a specific pathology such as fibroadenoma, it is clearly inappropriate for radiological *signs* of abnormality such as asymmetry. The problem can be circumvented to a limited extent by the use of synthetic or partially synthetic mammographic images, or of specimen X-rays. Fully synthetic images can either be created digitally using graphical techniques, or by imaging and/or digitising a synthetic object such as a radiographic phantom. Partially synthetic images can be created by combining synthetic abnormalities with real backgrounds, or vice versa. A synthetic abnormality can be created digitally by drawing lesion-like features into a digital image, or by extracting relevant image features from a real image. Films of normal and abnormal structures can also be physically superimposed prior to digitisation. In addition, it is possible to generate a synthetic abnormality prior to the imaging stage, by creating a lesion-like object and either embedding it within a mastectomy specimen or X-raying it alone. Purely synthetic images, including X-rays of mammographic phantoms and digitally created target-background pairs, are really only valuable in the early stages of developing and validating techniques.

We have approached this problem of establishing the ground truth in two ways, using pathologically proven examples, and radiological consensus. Both approaches have their drawbacks. We have used only biopsy-proven examples for developing detection and analysis methods for microcalcifications; however, it can be argued that this biases our data set towards difficult and obviously malignant examples, and away from those clusters which are perhaps very small, or obviously benign. In most of our work, we have chosen to use radiological consensus to define the ground truth, since this allows the identification of normal structures such as the gland disc, and it is based on mammographic appearance, that is, the appearance of those images we are attempting to interpret. Our consensus is generally obtained from independent annotations marked with a fine pen on registered acetate overlays. These are digitised, and may then be edited by the annotating radiologist. The simplest form of consensus involves

using only regions agreed upon by all of the annotating radiologists. This method has the advantage of eliminating spurious annotations, and the disadvantage that many subtle abnormalities will be excluded from analysis. A more sophisticated approach is to assign probabilities to regions depending on the number of radiologists in agreement about their status. A significant problem with the consensus approach has arisen in our work on breast asymmetry, where two radiologists were asked to delineate the glandular disc in a set of mammograms³⁶; this is a difficult, subjective task, and there were distinct variations in outline between the two sets of annotations. Our experiments were intended to compare the *shapes* of the left and right glandular discs; a simple combination of the two sets of annotations was thus inappropriate, as it generated features unrelated to the underlying structures. We eventually elected to use both sets of annotations independently.

It is important to determine the standard of performance at which we are aiming. It is likely that all our methods will be subject to error, so we must determine what error rates will be clinically acceptable. Acceptable levels of performance are dependent upon the task in hand; a pre-screening system is likely to have more stringent requirements than a prompting system, but both may ultimately achieve a similar cancer detection rate. We can determine acceptable standards for both prompting and classification systems by investigating the effects of different error rates on radiologists' performance using the systems. That said, it is unlikely that any system which cannot either perform at least as well as an expert radiologist, improve a less experienced observer's performance to that of an expert radiologist, or improve an expert radiologists' performance (perhaps reducing the necessity for double reading), will find clinical acceptance. Further issues include the degree of accuracy of any measurements required either for human purposes (such as assessment of therapy, and treatment planning) or for subsequent machine analysis and classification. This will also have a bearing on the required spatial resolution for analysis. For classification systems, the effect of errors both in diagnosis and in probability estimation must also be analysed.

Once the appropriate performance goal has been determined, extensive evaluation is essential if methods are to be accepted for clinical use. Receiver operating characteristic (ROC) analysis, supported by substantial trials on carefully selected data sets appear to be the methods of choice. Problems involve knowledge of the ground truth about the images used for evaluation (discussed above) and recruitment of sufficient experts both to provide opinions about the original films, and to participate in evaluation studies. It is also important to look carefully at the cases where automated methods fail, to assess the significance of such errors, and to identify any side-effects. For example, some methods for enhancing linear structure also have a broadening effect on peaks, which could make microcalcifications appear larger and less suspicious¹⁴. Technical assessment using ROC analysis should precede large scale evaluation by radiologists, both to avoid unnecessary exposure of experts to the test data set, and to maintain goodwill and optimism in the radiological community.

4.3 Develop computer-based methods for detection and analysis

We have developed computer-based methods for the detection of three major signs of mammographic abnormality; microcalcifications, asymmetry and spiculated lesions. There are a number of areas of similarity in our approach to detecting these different signs, including the use of radiologists' annotations to provide the ground truth for training and testing, training by example, the use of probabilistic methods, and the combination of evidence to increase detection performance. The work on analysis described in this paper is described in relation to clusters of microcalcifications, but the approach is also suitable for other forms of abnormality.

Detection of microcalcifications

Our approach to the detection of microcalcifications is based on grey-level mathematical morphology⁴⁰. Previously we have described a method by which cues for microcalcifications were generated by applying a morphological inner-edge detector (an eroded image subtracted from the original) and the top hat transformation³. This latter operator was formed by eroding the original image until all microcalcification-sized objects disappeared, and then dilating by the same amount to restore the background. All structuring elements used were approximately circular. Radiologists were asked to identify the locations of over 900 individual microcalcifications in a set of twenty image patches, thus identifying regions which were known to contain microcalcifications, and regions which were known to represent normal background. This enabled us to gather statistical distributions of both on- and off-target responses by our cue generators, which we then used to create images representing the likelihood of the presence of microcalcification at each image point, based on prior knowledge of the response of the cue generators. Using Bayesian statistics, we can combine evidence from different cue generators. Our results were presented as ROC curves; there was no overlap between training and test data, as a leave-one-out methodology was employed. These results demonstrated that the systematic combination of evidence can lead to improvements in detection performance².

We expressed our results in terms of individual microcalcifications detected rather than clusters detected, since it is a simple step to apply thresholds and clustering rules once individual particles have been identified correctly. A consequence of this was that our test set comprised a large number of very subtle particles, identified by radiologists studying both the original mammogram and a magnified digital version. Other published work focuses on the detection of clusters; most clusters of microcalcifications comprise both subtle and more readily detectable particles, so it is possible to achieve good cluster detection rates, whilst not actually detecting very subtle individual particles. We achieved a true positive pixel classification rate of over 97%, with a false positive rate of less than 2%.

Our method had a number of drawbacks, some of which we have since addressed. Firstly, we had problems training and testing our system, because of the difficulty in establishing a 'gold standard', that is, in determining exactly what each image contains. Our approach involved annotations made by an experienced radiologists who was

instructed to “mark all calcifications”. When we analysed our results in depth, we found our results degraded by mis-identifications in our training and test data. Initially, all annotations were made on registered acetate overlays, with subsequent revision on a computer monitor. This method can be improved by using multiple radiologists to annotate the images, allowing either the exclusion of regions of disagreement from subsequent analysis, or the assignment of probabilities based either on degree of confidence in the annotations or on number of radiologists in agreement. Another problem with our method arose because the top hat transformation as described above will detect all image peaks smaller than the defined maximum size. Study of the combined cue images has shown that many such peaks are below the size of genuine microcalcifications in our images, so we are now investigating pre-processing with a median filter or a single pass of erosion followed by dilation (‘closing’).

Detection of spiculated lesions

We have performed preliminary studies evaluating a method for detecting spiculated lesions by identifying characteristic patterns of co-radial lines³. This method is based on the Hough transform, and was tested on a set of twenty mammograms digitised with an effective pixel size of 0.2mm. A neighbourhood of equivalent diameter 2cm was defined, which determined the maximum diameter of any spiculated lesion detectable by the method. Neighbourhood size was limited by an assumption that only one spiculated lesion would reside in each such neighbourhood. All mammograms were pre-processed with Dixon’s line operator¹⁴. The neighbourhood was then systematically moved across each mammogram, with an overlap of 25% in width and height, selected as a compromise between execution time and sensitivity. At each position of the neighbourhood, the image was transformed in such a way that co-radial lines become co-linear points in transform space¹⁵.

The transform space can be represented as a grey level image. If there exists a single pattern of co-radial lines in the current neighbourhood of the line-enhanced original image, it will appear in the transformed image as a bright line. The gradient and intercept of this line define the centre of the co-radial pattern in the original image. We can thus detect the strongest co-radial pattern in the current neighbourhood by fitting a line to the data in the transform space, and we can then estimate the likelihood of the detected pattern representing a stellate lesion by measuring various properties of the line.

We have made three measures in transform space, after performing a least-squares line fit to the data. The first measure, the linear correlation co-efficient, characterises the degree of organisation of any lines in the enhanced original image. The second measure, designed to characterise the amount of line information related to the strongest pattern, is defined as the total number of ‘votes’ in a corridor along the fitted line. The third measure gives the degree of spread of lines around the focus of a detected co-radial pattern, by examining the distribution of evidence along the fitted line. All three measures were transformed back into the original image space, with values written at the detected foci. We thus produced three intrinsic images, each

characterising a different property. The data from the three images was again combined using a Bayesian approach².

Our results are encouraging, particularly for fatty breasts in which all the lesions in our data set were correctly identified. The experiments must, however, be regarded as very preliminary, with such restricted data for training and testing. Training was particularly problematic; attempts to model the distributions of measure responses failed, and we were forced to use smoothed versions of the raw distributions, which gave rise to inaccurate probability assignments.

Detection of asymmetry

Asymmetry between contra-lateral breast images is an important indicator of breast disease. However, it is difficult to detect asymmetry automatically, as radiologists appear to use a sophisticated analysis procedure incorporating radiological, anatomical and pathological knowledge. We have performed experiments to elucidate the type and degree of asymmetry considered significant by radiologists. Results of experiments in which radiologists were shown only the boundaries of non-fatty regions in mammograms indicate that, in addition to densities, radiologists may use shape, size, location or topology of such regions in their assessment, as they were able to achieve approximately 70% accuracy in discriminating significant asymmetries from benign asymmetries and normal mammograms³⁶.

Our approach to the detection of asymmetry between left and right breast images attempts to utilise this result. We first segment each breast into regions of like tissue type³⁵; this enables both an analysis of the shapes of individual regions, and direct comparison of similar breast structures. Our best segmentation results to date, using one of Laws' texture energy measurements³², achieve over 80% agreement with radiologists. The second stage of our approach involves measuring and comparing shape and grey-level properties of the detected regions. A variety of methods have been investigated, with the best classification accuracy of nearly 87% being obtained from a statistical combination of measures. One difficulty with shape measurement, however, is the fact that many simple shape measures are rotationally invariant; this does not fit well with our knowledge of breast anatomy.

Registration and normalisation remain areas of difficulty for comparing contra-lateral breasts and for comparing successive examinations of the same breast. Further psychological experiments are required to determine, for example, whether the absolute size of the non-fatty region is more significant than the size relative to the breast in which it is sited, and to what degree a difference in size between the two breasts affects radiologists' judgements about asymmetry. The results will have important implications for automated analysis. For those methods which require registration, we have elected to avoid distortion-based methods, on the grounds that they are likely to modify significant image structure.

Analysis of microcalcification clusters

We are investigating a new, low cost approach to the identification and delineation of mammographic abnormalities for training and test purposes. Although our methods

are described in terms of clusters of microcalcifications, the overall strategy is also applicable to other localised abnormalities. For this work we have elected to use X-rays of excised lesions. Whilst we realise the importance of working with the images we are ultimately intending to interpret, the use of images of excised abnormalities has a number of advantages including superior image quality, reduced obscuration by overlying tissue, known pathology, and the freedom to experiment with imaging parameters. Such images provide a valuable half-way-house between the analysis of images of synthetic lesions and of images of lesions within patients' breasts.

Our method extracts morphologic descriptions of excised clusters of microcalcification from multiple X-ray views. A specimen of tissue containing a cluster is chilled to increase its rigidity, and secured inside a small cardboard Tetrahedron edged with fuse wire. The edges of the Tetrahedron provide a reference frame so that, when the Tetrahedron is X-rayed lying on each of its faces, features in the four resulting images can be related. At present we are working with three types of specimen providing images of varying degrees of difficulty: X-rays of fragments of lead embedded in normal breast tissue from a reduction mammoplasty; acetate overlays of X-rays of genuine microcalcification clusters; and X-rays of genuine microcalcification clusters.

The first stage of our method involves segmenting the images. This is performed semi-automatically, using histogram-based thresholding to extract particle pixels from digitised acetates, and a line detection algorithm¹⁴ to extract the fuse wire triangle at the base of the Tetrahedron in the current view. Microcalcifications are identified manually in the X-rays of genuine clusters, as our automated detection method is insufficiently specific. The accurate identification of the reference frame is crucial. Any spurious lines are eliminated by a process of edge-linking, and by the constraint that the three lines we are seeking should have 60° angular separation. Straight lines are fitted to the resulting data, and the centre of gravity of the triangle can then be identified. Any spurious particles are removed automatically on the basis of size, or manually. The procedure is repeated for all four views.

The boundary of each detected particle is stored as a list of real co-ordinate pairs. An error vector, normal to the local boundary curve, is associated with each point; this allows compensation for tissue mobility, inaccurate Tetrahedron construction and the imaging and detection processes. The approach used for identifying common features in different views is based on the 'auxiliary line' method described by MacKay³³. This method uses the fact that an object which appears as a single point in one X-ray projection could appear anywhere along a line in another projection. The gradient and intercept of the line depend on the relative projections. In our case, we have an object which appears as a core region (a calcification particle) surrounded by an error region defined by the error vectors. In a second view, the object will appear on a band, of which the limits are defined by the extremities of the particle in the first view. The band, which we call the 'auxiliary band', is surrounded by an auxiliary error region defined by the limits of the error region in the first view. For each microcalcification in the first view, we seek a match along the appropriate auxiliary band in the other three views. In some cases the match is unambiguous; there is only one calcification within the auxiliary band. In other cases, there may be multiple candidates; these are resolved

by seeking confirmation from other projections, and by matching basic properties such as total density.

To date, we have successfully matched 'calcifications' from the images of lead shot embedded in real breast tissue, and from acetate overlays of clusters containing up to approximately thirty particles. Clearly, in the latter case, we have neither particle shape nor density to aid the matching process. A significant difficulty with both the acetate images and with the specimen X-rays is the problem of missing data. In many cases, we found particles identified (or identifiable) in only one projection, either because of superposition of tissue, superposition of particles, or because the particle was very small, but elongated and could only be distinguished when it was aligned perpendicularly to the X-ray plane.

The potential uses of this technique include three-dimensional classification of cluster shapes to validate the two-dimensional approach of Lanyi³⁰, and the extraction of detailed, quantitative descriptions of the radiological appearance of abnormalities with known pathology, from high quality images. Once a high-resolution, three-dimensional representation of an abnormality has been formed, it can be rotated to resemble the appearance in the original mammogram, allowing more accurate identification of normal and abnormal structures for training and test purposes.

4.4 Find out how these methods can be used effectively

We have described a number of computer-based methods for the detection and analysis of breast abnormalities, but we have not yet addressed the issue of how such methods might be used to improve performance in clinical practice. Since pre-screening is some distance away, we have focussed our efforts on prompting, in which we indicate the locations of automatically-detected suspicious regions to radiologists. As indicated earlier, we believe that there are many unanswered questions associated with prompting. We have started to investigate these areas by studying the effects of prompting radiologists with the locations of clusters of microcalcifications detected using a method based on that described above²⁴.

Our subjects, five experienced breast radiologists, were asked to locate microcalcification clusters in digital images presented with and without prompts. Our data set consisted of a set of 30 mammograms, half of which were normal and half of which had been classified as abnormal on the basis of a single microcalcification cluster. From each of these mammograms, a central 15cm square region was extracted and digitised. The resulting digital image was 1024x1024 pixels in size, and had a spatial of 0.15mm pixel⁻¹, with 8 bit grey resolution. In each case, the digitised region was selected to include as much of the breast tissue as possible; in abnormal images, the microcalcification cluster was always present in the digitised region. The digital images were displayed on a sun SPARC workstation with a pixel size of 0.3mm.

Each of the 30 images was processed by an automatic cluster detection system based on the method described earlier. For each region identified as a potential cluster by the system, a prompt was generated. Prompts took the form of open red circles, 100

pixels in diameter, superimposed on the digital image. The images were then presented, in both processed and unprocessed forms, to each of the radiologists. The order of the 60 presentations was randomised differently for each subject, with the constraint that two occurrences of the same image were separated by at least 20 presentations.

Prior to the experimental session, each subject was given a standardised set of verbal instructions and 6 practice images with which to familiarise themselves with the task and the equipment. The practice images included examples of each of the experimental conditions but did not include any images that were used in the main experiment. The verbal instructions contained the following information: the number of presentations; the requirements of the task; the ratio of normal/abnormal images; the ratio of processed/unprocessed images; the approximate accuracy of the prompt generation algorithm.

For each presentation, the subjects were required to study the image, identify any potential clusters and mark their locations by means of a cursor controlled by a mouse. After marking each suspicious location, the subjects were required to indicate their confidence that the marked location represented a cluster by means of the following five point scale: definitely a cluster; probably a cluster; possibly a cluster; possibly not a cluster; probably not a cluster. A sixth scale point, 'definitely not a cluster', was represented by a 'next image' option, which the subjects could select without making a location judgement if they believed that there were no clusters in the image. The subjects were asked to keep their interpretations of the confidence levels consistent throughout the experiment, and to try to use the whole scale.

The subjects were able to make as many location judgements as they thought appropriate on any given image, with each location judgement followed by a confidence judgement. Once the subjects had finished making location and confidence judgements, or if they wished to make no such judgement, they were able to move on to the next presentation by selecting 'next image' with the mouse. This allowed the subjects to control the time spent studying each image, which reflects the normal film screening environment better than a fixed presentation time. In addition, the study time was recorded for further analysis. In those presentations where prompts were available, they were briefly displayed for 200msec when the image first appeared. This should have served to direct the attention of the subjects to the prompted region and alert them to the availability of the prompts, while minimising the effect of the prompts as distractors. After this initial brief presentation, the subjects were allowed to toggle the prompts on and off as they desired.

The performance data from each subject were recorded and processed by means of receiver operating characteristic (ROC) analysis. The detection sensitivity of the radiologists, in terms of the signal detection measure d' , was observed to be significantly higher than that of the automated detection system, both when the radiologists were prompted, ($t_{\text{obs}} = 13.22$, $p < 0.0005$), and when they were unaided, ($t_{\text{obs}} = 3.33$, $p < 0.025$). Figure 1 shows the composite ROC curves for the subjects in both the prompted and unprompted conditions. The most important result of our study is that the detection performance of the radiologists was significantly higher

when the images were presented with prompts than when they were unprocessed, ($t_{\text{obs}} = 3.47$, $p < 0.025$). Our results are consistent with the findings of Chan, who suggested that an observer working in conjunction with a computer-aided diagnosis (CAD) system is more effective than either the observer or the CAD system working alone.

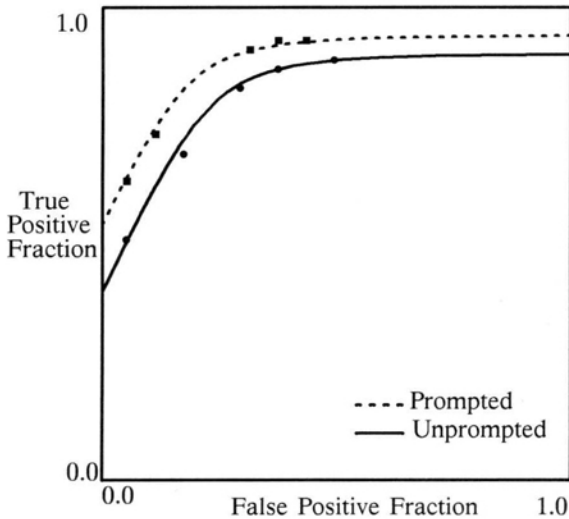


Figure 1. ROC curves showing improvements in radiologists' detection of microcalcification clusters using prompting compared with their performance without prompts.

Having determined that prompting had a significant effect on detection performance, we analysed the ways in which prompts were used by the subjects. Prompts were available on 26 of the 60 images, and after the initial brief display, subjects were able to switch them on and off as desired. Therefore there was a distinction between the initial passive display and any subsequent active use of the prompts. Overall, active use was made of the prompts in 34% of the cases where they were available, with individual subjects varying between 15% and 69%. The cases in which prompts were most frequently used, that is, those images on which two or more subjects made active use of the prompts, generally corresponded to those cases in which more than one prompt was displayed. In 13 of the 15 images with more than one associated prompt, the prompts were actively used by more than one subject, while active use by two or more subjects occurred in only 2 of the 11 images that had only a single associated prompt. The correlation between the number of prompts associated with an image and the number of subjects who actively used the prompts on that image was calculated to

be 0.61. The generation of multiple prompts is a likely result of any CAD system. Our results show one possible consequence of having multiple prompts associated with an image: the need to examine the prompt information in more detail. It seems as though, as would be expected, a single prompt acts to cue attention, directing it towards the prompted region, in which case a single brief presentation is sufficient. However, when multiple prompts are presented, the indivisibility of attention requires that each be checked in turn. It is possible that the need to check each prompt in sequence may impose artificial constraints on the natural search pattern of the observer, requiring a greater level of cognitive processing and consequently an increase in the time required to study the image.

An analysis of the study time per image revealed that the radiologists took significantly longer to examine the processed images than the unprocessed images, ($t_{\text{obs}} = 2.64$, $p < 0.025$). However, the subjects were not aware that their study times were being recorded during the experiment and were therefore not under any pressure to respond rapidly. This may make the study time data less reliable than the accuracy data.

Our results have demonstrated that a prompt generation system with a moderate false-positive rate be of some benefit to the radiologist screening for clustered microcalcifications, at least in an experimental setting. There are a number of issues that still need to be addressed before prompting could be effectively implemented in a clinical environment and we are currently undertaking experiments to investigate the effects of varying the accuracy of the prompt generation system, and the effect of prompting on the detection of non-prompted abnormalities.

5. CONCLUDING REMARKS

Considerable effort has been invested in developing methods for automatically detecting mammographic signs of breast disease. The greatest success has been in the detection of microcalcifications, which are relatively well-defined and thus amenable to detection using conventional computer vision methods. We have described a framework into which such methods could be placed. Five important areas which require attention before such a system can be realised are; identifying the most appropriate targets for assistance, determining performance standards for different tasks, defining a mechanism for demonstrating clinical acceptability, developing new techniques for the automatic detection and analysis of signs of abnormality, and determining the effects of computer-based aids on human performance.

We have described progress in Manchester towards these goals. In particular, we have described: experiments to investigate the factors which affect radiologists' ability to detect microcalcifications; methods for detecting microcalcifications, spiculated lesions and asymmetry; a new approach to extracting quantitative descriptions of mammographic abnormalities from biopsy specimens; and a demonstration that computer-generated prompts can significantly improve radiologists' performance in detecting microcalcification clusters.

ACKNOWLEDGEMENTS

This research has been funded primarily by the Cancer Research Campaign, with additional support from the Science and Engineering Research Council, the Wolfson Breast Pathology Unit and IBM. We are grateful to Prof Alastair Gale of the University of Derby for many helpful discussions, and to the radiologists, radiographers and physicists who have given their time to assist with our experiments.

REFERENCES

- 1.S. J. Adamson "Perceptual recognition of breast cancers." *MSc Thesis, University of Manchester*, 1992.
- 2.S. M. Astley and C. J. Taylor "Combining cues for mammographic abnormalities." *Proc. 1st British Machine Vision Conference*, pp. 253-258, 1990.
- 3.S. M. Astley, C. J. Taylor, C. R. M. Boggis, D. L. Asbury, M. Wilson "Cue generation and combination for mammographic screening." Chapter 13 of *Visual Search II*, ed. D. Brogan, Taylor and Francis, London, 1992.
- 4.C. Bourrely and S. Muller "Detection of microcalcifications in mammographic images." *Neurocomputing (NATO ASI Series, Vol. F68)* eds. F. Fogelman Soulie and J. Herault, Springer Verlag, Berlin 1990.
- 5.D. Brzakovic, X. M. Luo and P. Brzakovic "An approach to automated detection of tumours in mammograms" *IEEE MI-9*, Vol. 3, pp. 232 - 241, 1990.
- 6.P. J. Burt "The pyramid as a structure for efficient computation." *Multiresolution image processing and analysis*, ed. A. Rosenfeld, pp. 6-36, Springer-Verlag, 1984.
- 7.C. B. Caldwell et al "Characterisation of mammographic parenchymal pattern by fractal dimension." *Physics in Medicine and Biology*, Vol. 35, No. 2, pp. 235-247, 1990.
- 8.H-P. Chan et al "Evaluation of unsharp mask filtering for the detection of subtle mammographic microcalcifications." *Proc. SPIE International Society of Optical Engineers (USA)*, Vol. 626, Pt. 1, pp. 347-348, 1986.
- 9.H-P. Chan et al "Computer-aided detection of microcalcifications in mammograms." *Investigative Radiology*, Vol. 23, No 9, pp. 664-671, 1988.
- 10.H-P. Chan et al "Improvement in radiologists' detection of clustered microcalcifications in mammograms: the potential of computer aided diagnosis." *Investigative Radiology*, Vol. 25, pp. 1102-1110, 1990.
- 11.A. R. Cowen et al "The computer enhancement of digital grey-scale fluorography images." *British Journal of Radiology* Vol. 61, pp. 492-500, 1988.
- 12.D. H. Davies and D. R. Dance "Automatic computer detection of clustered microcalcifications in digital mammograms." *Physics in Medicine and Biology*, Vol. 35, No. 8, pp. 1111-1118, 1990.
- 13.A. P. Dhawan and E. Le Royer "Mammographic feature enhancement by computerized image processing." *Computer Methods and Programs in Biomedicine*, Vol. 27, pp. 23-35, 1988.
- 14.R. N. Dixon and C. J. Taylor "Automated asbestos fibre counting." Chapter 4 of *Machine Aided Image Analysis*, ed. W. E. Gardner, Institute of Physics Conference Series No. 4, 1979.

15. T. P. Ellison "Detection of Stellate Lesions in Mammograms" *MSc. Thesis*, University of Manchester 1989.
16. M. Flynn et al "Replication of diagnostic radiographs using a film scanning/printing system." *SPIE Medical Imaging IV: Image Capture and Display*, Vol. 1232, pp. 88–96, 1990.
17. A. P. M. Forrest "Screening for breast cancer: the UK scene." *British Journal of Radiology*, Vol. 62, pp. 695–704, 1989.
18. A. P. M. Forrest and R. J. Aitken "Mammography screening for breast cancer." *Annual Reviews of Medicine*, Vol. 41, pp. 117–132, 1990.
19. A. G. Gale, G. E. Walker, E. J. Roebuck and B. S. Worthington "The quest for accuracy, consistency and uniformity of performance in mammographic screening: the systematic imperative." *British Journal of Radiology* Vol. 62, S10, 1989.
20. M. L. Giger et al "Image features of mammographic masses used in the development of computerized schemes." *SCAR 90: Computer Applications to Assist Radiology*, eds. R. Arenson and R. M. Friedenber, 1990.
21. R. Gilles, F. Bouvet-Lefebvre, J. Masselot and E. Kahn "Characterisation of benign clustered microcalcifications with a new image analysis method" *Proceedings of Computer Assisted Radiology*, 1991.
22. HMSO *Breast Cancer Screening (the Forrest Report)*, Her Majesty's Stationery Office, London, 1986.
23. A. Hoyer and W. Spiesberger "Computerized mammogram processing." *Philips Technical Review* Vol. 38, Part 11/12, pp. 347–355, 1978/79.
24. I. W. Hutt "The effects of prompting on the detection of microcalcification clusters in digital mammograms." *MSc Thesis, University of Manchester*, 1992.
25. N. Karssemeijer "A stochastic method for the automated detection of microcalcifications in digital mammograms." *Proc. XIIth Conference on Information Processing in Medical Imaging, Wye College, Kent*, 1991.
26. W. P. Kegelmeyer "Computer detection of stellate lesions in mammograms" *SPIE Biomedical Image Processing and Three-Dimensional Microscopy*, Vol. 1660, 1992.
27. C. Kimme, B. J. O'Loughlin, J. Sklansky "Automatic detection of suspicious abnormalities in breast radiographs." *Data Structures, Computer Graphics and Pattern Recognition*, ed. Klinger, pp. 427–447, Academic Press, New York, 1975.
28. H. L. Kundel and C. F. Nodine "Studies of eye movements and visual search in radiology" In: *Eye movements and the higher psychological function* ed. J. A. W. Senders, D. Fisher and R. Monty, Hillsdale NJ, Lawrence Earlbaum Associates, 1978.
29. S. M. Lai, X. Li and W. F. Bischof "Automated detection of breast tumours" *Computer Vision and Shape Recognition*, eds. A. Krzyzkak, T. Kasvand and C. Y. Suen, World Scientific Series in Computer Science, Vol. 14, pp. 115–132, World Scientific, Singapore, 1989.
30. Lanyi M. "Morphological analysis of microcalcifications." In: *Early Breast Cancer*, ed J. Zander and J. Baltzer, Springer-Verlag, Berlin, 1985.
31. T-K. Lau and W. Bischof "Automated detection of breast tumors using the asymmetry approach." *Computers and Biomedical Research*, Vol. 24, pp. 273–295, 1991.

32. K. I. Laws "Textured image segmentation" *Image Processing Institute Report*, No. 940, University of Southern California, Los Angeles, 1980.
33. S. A. MacKay, M. J. Potel and J. M. Rubin "Graphics methods for tracking three-dimensional heart wall motion." *Computers and Biomedical Research*, Vol. 15, pp. 455-473, 1982.
34. I. E. Magnin, F. Cluzeau, C. L. Odet "Mammographic texture analysis: an evaluation of risk for developing breast cancer." *Optical Engineering*, Vol. 25(6), pp. 780-784, 1986.
35. P. I. Miller and S. M. Astley "Classification of breast tissue by texture analysis." *Image and Vision Computing*, Vol. 10, No. 5, pp. 277-282, 1992.
36. P. I. Miller and S. M. Astley "Detection of Breast Asymmetry Using Anatomical Features" *Proceedings of SPIE Biomedical Image Processing IV, San Jose*, 1993
37. P. H. M. Peeters, A. L. M. Verbeek, J. H. C. L. Hendriks, R. Holland, M. Mravunac and G. P. Vooijs "The occurrence of interval cancers in the Nijmegen screening programme." *British Journal of Cancer* Vol. 59, pp. 929 - 932, 1989.
38. The Pritchard Committee "Guidelines on the establishment of a quality assurance system for the radiological aspects of mammography used for breast screening." *UK Radiation Advisory Committee: Subcommittee on Quality Assurance*, 1988.
39. J. L. Semmlow, A. Shadagopan, L. V. Ackerman, W. Hand and F. S. Alcorn, "A fully automated system for screening xeromammograms." *Computers and Biomedical Research* Vol. 13, pp. 350-362, 1980.
40. J. Serra *Image Analysis and Mathematical Morphology*, Academic Press, London, 1982.
41. A. Shadagopan, F. S. Alcorn, J. L. Semmlow, L. V. Ackerman "Computerized quantification of breast duct patterns." *Radiology*, Vol. 143, pp. 675-678, 1982.
42. E. A. Sickles "Mammographic evaluation of breast calcifications" *Radiology* Vol. 160, pp. 289-293, 1986.
43. A. Treisman "Preattentive processing in vision." *Computer Vision, Graphics and Image Processing*, Vol. 31, pp. 156-177, 1988.
44. J. N. Wolfe "Risk for breast cancer development determined by mammographic parenchymal pattern." *Cancer*, Vol. 37, pp. 2486-2492, 1976.