The Computer-Aided Detection of Abnormalities in Digital Mammograms

A thesis submitted to the University of Manchester for the degree of PhD in the Faculty of Medicine

Ian Hutt

Department of Medical Biophysics

1996

Contents

Abstract

The most effective way to control Breast Cancer is to detect the early signs of the disease and treat it before it can develop into a more serious problem. In order to achieve this the UK National Health Service operates a screening programme for all women between the ages of 50 and 65. The screening programme uses mammography, which is the most suitable of the available techniques for imaging the breast. However, the effectiveness of mammography critically depends on the ability of the radiologist to detect the small, often very subtle abnormalities that may be present in a mammogram. Prompting is a technique designed to aid radiologists in the detection of the subtle signs of breast cancer by directing attention towards the potentially suspicious regions of a mammogram that have been identified by computer–based detection systems. Previous research has suggested that a radiologist working in conjunction with a sufficiently accurate prompting system can lead to improvements in detection sensitivity.

The aim of this thesis is to investigate the feasibility of prompting as an aid to the radiologist. This includes an investigation of the computer–based techniques that may be used to automatically generate prompts and a study of errors that such a prompting system might make. Errors made in generating prompts may affect the search strategy and detection performance of the radiologist. Experimental studies described in this thesis have been used to investigate the effects that errors in prompt generation have on the search performance of the radiologist. This work suggests that prompting can be an effective technique for aiding the radiologist in the early detection of Breast Cancer, providing that certain conditions concerning the accuracy of prompts are met.

Declaration

No portion of work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright Notes

Chapter 1

Introduction and Background

This chapter is designed to provide a general background on breast cancer, the importance of mass screening in controlling the disease and some techniques that are used for diagnosis, particularly mammography. The final section provides an overview of the remainder of the thesis and explains the purpose of this work.

1.1 Breast Cancer

Each year in Britain there are 24500 new cases of breast cancer and 15000 deaths from the disease. Around 20% of all new female cancer cases in Britain involve breast cancer, making it the most common form of cancer among women. It has been estimated that one in every twelve women will be affected by the disease at some time in their lives (Asbury 1990).

These rather bleak statistics illustrate the scale of the breast cancer problem, a problem that is compounded by the fact that although a number of risk factors have been identified, there are no known primary preventative measures for breast cancer and no specifically directed cures (Strax 1981).

There are a number of factors suggested as being associated with an increased probability of developing breast cancer. These risk factors include early onset of

menarche, late onset of menopause, nulliparity, late age at first child birth, a history of benign breast disease, hormone replacement therapy and certain dietary factors (Henderson et al 1984). In particular, a family history of breast cancer is associated with an increased risk of developing the disease. The risk of disease may double or triple for first degree relatives of breast cancer patients (Anderson 1974), and can even rise as high as a nine times increase for the first degree relatives of premenopausal women with bilateral breast cancer (Henderson et al 1984).

However, in over 75% of women with breast cancer none of these risk factors are present (McClelland 1990). Furthermore, Berg (1984) has suggested that the only clearly identifiable risks associated with the disease are gender and ageing.

The relationship between ageing and the probability of developing breast cancer is clear. Roebuck (1990) investigated the ages of 1500 women in one institution who were found to have developed breast cancer. The distribution by age of these women is illustrated in figure 1.1 which clearly shows the increased chance of developing the disease after the age of 40. Other studies have shown very similar patterns (Macmahon et al 1973, Henderson et al 1984, Austoker et al 1988).

The lack of any risk factor, other than ageing, that is strongly predictive of disease occurrence makes it difficult to develop any measures for the primary prevention of breast cancer.

Breast cancer is an extremely complex condition that may assume a variety of forms and several models have been proposed that describe the clinical development of the disease (Gallagher 1985). Carcinomas generally arise either from the cells in the lobules of the milk–secreting system or in the branching duct system that transports milk to the nipple. They are subdivided into non–invasive and invasive types depending on whether the malignant cells are entirely confined within the lobules/ducts, or whether they have spread into surrounding tissue.

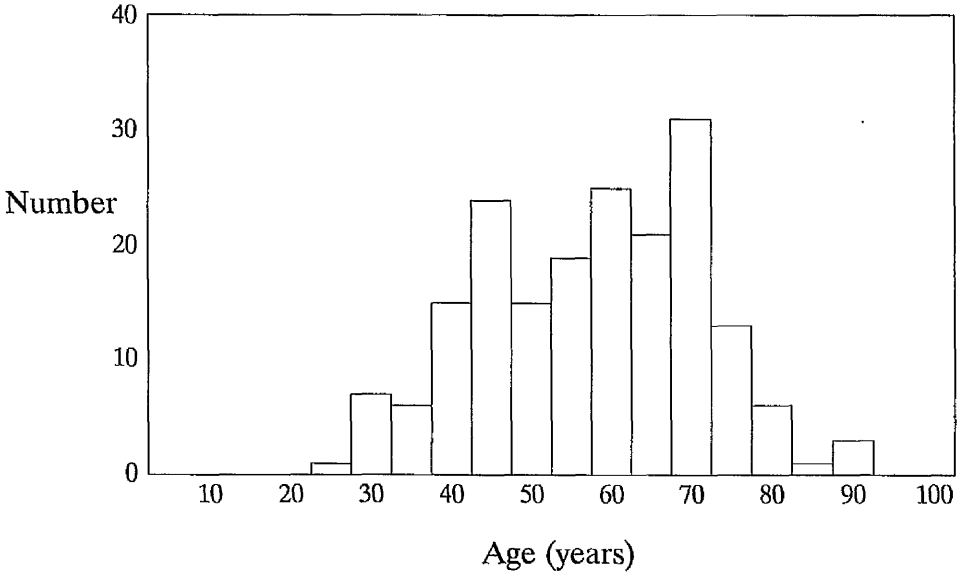

Figure 1.1: Distribution of incidence of breast cancer by age.

While a detailed discussion of the suggested pathogenic courses of breast cancer is beyond the scope of this report, it should be noted that despite the large amount of research that has been carried out in this area, the cause or causes of the disease are still not fully understood (Baum 1988). However, the initial clinical manifestation of breast cancer is generally observed to be a single localised lesion in one breast. This stage in the development of the disease is commonly referred to as "early" or "minimal" breast cancer.

There is no generally accepted definition of precisely what constitutes an early cancer. Gallagher and Martin (1971) refer to minimal cancer in cases when the tumour is no larger than 0.5cm in size, while Urban (1976) raises this size limit to 1.0cm, with the additional stipulation that lymph node metastases cannot be palpated. Even the fact that a carcinoma is clinically occult does not necessarily lead to its classification as early cancer, since it is possible that relatively large tumours may remain clinically occult in large breasts (Lanyi 1985). In addition, Fisher (1985) points out that a small (0.5cm) carcinoma, although regarded as

early in the clinical sense, represents a cell mass that has undergone 27 population doubles and is therefore, biologically speaking, a late tumour.

In view of this lack of a precise definition, for the remainder of this text the term "early cancer" will be used to refer to any small (less than 1.0cm), localised, non–invasive lesion arising from an early stage in the development of breast cancer. It is recognised that at this stage of the disease there may be some microscopic, occult dissemination elsewhere in the breast tissue but, as pointed out by Strax (1981), it is generally accepted that an intact immunological system can cope with this if the main clinical tumour burden is removed.

The concept of early cancer seems to suggest that the key to the successful treatment of breast cancer may lie in the detection and treatment of the localised lesion at a stage when the body's natural immunological system is still intact. In fact there is a substantial body of evidence to support this notion of early detection and treatment. For example, Figure 1.2 shows the survival rates of patients up to 5 years after the diagnosis and treatment of breast cancer, depending on the stage of development of the disease at diagnosis (CRC Factsheet 6 1988). Stage I corresponds to early breast cancer as described above. Stage II involves slightly larger tumours (2–5cm) with or without lymph node involvement. Stage III refers to locally advanced tumours possibly attached to the chest wall. At stage IV distant metastases are present.

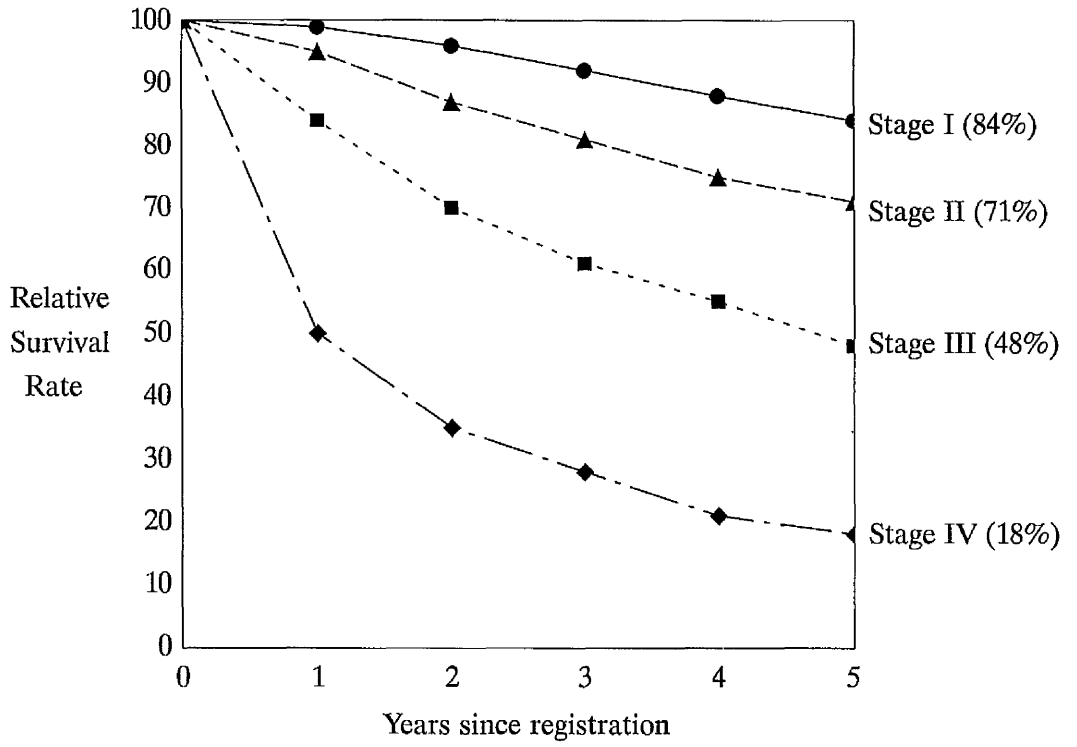

Figure 1.2: Breast cancer relative survival according to stage of development

In another study, Letton and his colleagues (1977) examined 5810 asymptomatic women between the ages of 35 and 50 and discovered that 32 of this group had breast cancer in the early stages of development; this was subsequently treated. They report a five year cure rate among these patients of 87.1% compared to the 63% rate commonly observed among unscreened women. Strax (1981) claims similar findings, reporting a five year survival rate of 85% in those cases where the cancer was treated at an early, localised stage. In those cases where treatment occurred at a later stage of development, after the glands had become affected, the five year survival rate dropped to 53%. In addition, Helman (1977) has reported a 10–15% reduction in the mortality rate for women over 50 who were screened using mammography, while Moskowitz (1977) has suggested that the early detection and treatment of small breast cancers may increase the twenty year survival rate from 47% to 70% among women affected by the disease.

These figures clearly demonstrate that the treatment of breast cancer at an early stage in its development can considerably improve the patient's chance of survival. An additional advantage of early detection is that recent advances in breast–conserving surgery and radiation therapy as an alternative to radical mastectomy (removal of the breast and associated structures) may mean that treatment of the disease at an early stage, while the tumour is still small, can greatly improve the possibility of conserving the breast (Haffty et al 1991).

However, the effectiveness of early treatment relies on the detection of the carcinoma at an early stage in its development, prior to the occurrence of any obvious symptoms that would generally indicate that the disease has progressed to a more advanced stage (Strax 1981). This means that women must be examined for breast cancer at a point before the presence of the disease is suspected by either the patient or the clinician.

1.2 Mammography

1.2.1 What is Mammography ?

Mammography is an X–ray technique for studying the breast. In common with other methods of radiographic examination the X–ray beam is passed through the breast and is differentially absorbed by the various types of tissue encountered. The emergent beam is then recorded as an image, or mammogram, on a sensitive film. The differential absorption of x–rays by different types of tissue means that the resulting mammogram represents a picture of the internal structure of the breast. Typically, radiopaque areas of the breast such as glandular and fibrous tissue appear as relatively bright areas in the mammogram, while the surrounding fatty tissue, which is radiolucent, appears darker. Since the various abnormalities that may be present in the breast all have their own particular absorption characteristics it is possible for the radiologist to identify potentially suspicious regions of the mammogram.

A full mammographic study consists of four films, with two views of each breast. These consist of a mediolateral oblique (side) view and a craniocaudal (top–down) view. However, in some centres it is current practice in screening mammography to use only a single, mediolateral view of each breast. The single view approach is generally justified on the grounds of reduced cost, reduced acquisition time (and therefore increased patient throughput) and reduced interpretation time (Tabar et al. 1983).

Some doubt has been expressed on the adequacy of the single view approach, primarily on the grounds that a single view is not sufficient to identify all mammographically detectable cancers. In addition, single view screening is more likely than the double view approach to require additional images to clarify potential abnormalities (Muir et al 1984). For example, Sickles et al. (1986a) studied 2500 asymptomatic women undergoing screening for the first time. The mammograms from each case were interpreted twice; once with only the single, oblique view and once with both the oblique and craniocaudal views. Figure 1.3 summarises the results.

	One View per breast	Two views per breast
Abnormal interpretations	642	179
Mammography–generated biopsies	76	83
Mammography–detected cancers	25	27

Figure 1.3: Results of baseline screening of 2500 women (Sickles et al 1986a)

In addition to the improved sensitivity displayed by double view mammography in this study, high extra costs were generated by the greatly increased number of follow–up images required for abnormal interpretations in the single view cases. These extra costs more than offset the savings made by using only a single view

– making the single view approach more expensive, as well as causing unnecessary anxiety in women who were recalled due to these false–positive interpretations (Sickles 1990a).

Results such as these have recently led the NHS to change its screening practice in favour of double view mammography.

During the imaging process the patient's breast is compressed against the radio–sensitive plate. Although this results in discomfort for the patient it is necessary in order to obtain an image of sufficient quality. Compression of the breast serves to minimise the effect of scattered radiation which can greatly reduce the contrast of the resulting film. It also enables lower doses of radiation to be used and reduces artifacts that may occur as a consequence of motion of the breast.

Apart from contrast, the other important measure of image quality is resolution and the most important factor in producing a sufficiently high resolution is the film/screen combination, together with suitable film processing. The processing of the films is in fact critical to obtaining a mammogram of suitable technical quality. Roebuck (1990) suggests that "more potentially good mammograms are ruined, often to the level of being non–diagnostic, by lack of care in the choice and control of processor chemistry, operating temperature and development time than by any other single factor".

Once a film of suitable technical quality is produced it must be viewed by a radiologist in order to locate any potentially abnormal structures. Clearly, it is no use producing technically good mammograms if the viewing conditions are inadequate to allow perception of all of the image detail in the film. Due to the sensitivity of the human visual system to small changes in brightness when the overall brightness is high, mammograms are generally viewed on high intensity film illuminators (light walls). Again there is a need for compromise, since at high

levels of intensity any unmasked portions of the illuminator may produce glare, which could adversely affect film reading performance (Roebuck 1990).

1.2.2 Why use Mammography ?

Probably the most straightforward method for examining the breast is by palpation, which may be performed either by the clinician or by the patient herself through self examination. However, there are certain limitations on the effectiveness of palpation as a method for detecting the signs of early breast cancer.

Firstly, the breast is a naturally multi–nodular organ which may make it difficult to distinguish the small nodule that indicates an early cancer from naturally occurring lumps in the breast. In addition a very small lesion may not be palpable even by the most expert clinician and yet may still represent the early stages of a potentially fatal disease. It is well known that a small tumour may be present in the breast while being completely asymptomatic and non–palpable. Clearly, in such cases it is necessary to use a further, more sophisticated method of examining the internal structure of the breast.

There is some evidence to suggest that ultrasound, or sonomammography, provides better estimates of tumour size than either mammographic or clinical examination (Fornage et al 1987). However, very small lesions may not be reliably detected by sonomammography, so although the technique may be useful when used in conjunction with mammography it cannot really be considered suitable as the sole method of investigation except in exceptional circumstances (Roebuck 1990). For example, an ultrasound examination may be appropriate when it is particularly important that the patient avoid ionising radiation – such as when the patient is very young or pregnant.

For many locations in the body, Computed Tomography (CT) scanning is able to distinguish much smaller differences in density than conventional x–rays,

providing useful additional diagnostic information. However, this improved density discrimination is less useful in examining the breast as mammography itself produces very high contrast images (Kopans 1987). Radiation doses during CT are considerably higher than those involved in mammography, the cost of a CT examination is very high and an intravenous infusion of iodide is required to allow reliable discrimination between benign and malignant structures (Sickles 1990b). In addition, the spatial resolution of CT is poorer, making the technique less reliable for detecting microcalcifications.

These factors mean that although CT scanning is inappropriate for the detection and diagnosis of breast cancer, the technique can be used in pre-biopsy localisation of lesions located very near the chest wall when they are difficult to image using mammography (Muller et al 1983).

Magnetic resonance (MR) imaging has been shown to image areas of dense fibro-glandular tissue with a greater contrast range than either mammography or CT scanning (Kopans 1987). However, the spatial resolution of MR imaging is inferior to that of mammography - making the detection of smaller lesions unreliable (Sickles 1990b). In addition, the inability of MR imaging to detect calcium-containing structures means that microcalcifications, an important sign of early breast cancer, cannot reliably be detected by this technique (Turner et al 1988).

As with CT scanning, these factors in combination with the high cost of examination make MR imaging inappropriate as a method of screening for breast cancer. However, MR imaging has been used as a complement to mammography for examining already detected lesions (Sickles et al 1990b).

Though it does not involve imaging, fine needle aspiration (FNA) cytology is a diagnostic technique commonly used in conjunction with mammography. A needle is inserted into an abnormal region in the breast and a sample of cellular material is extracted for pathological examination. Although this is a useful

diagnostic tool, and has some therapeutic value in the treatment of cysts, it does require that the abnormality has been detected before it can be used.

In summary, mammography is capable of providing images with high contrast and high spatial resolution at a relatively low radiation dose and a relatively low cost. The various problems associated with other imaging techniques make them less suitable for the detection of breast cancers, which is why mammography is currently the primary method used for breast screening.

1.2.3 The Appearance of Abnormalities in Mammograms

There are a number of different classes of abnormality that may be observed in mammograms and within any given class the appearance may vary greatly.

One of the most significant types of mammographic abnormality is microcalcification. Lanyi (1985) has described microcalcifications as "the most important leading symptom in mammographic detection of pre-clinical carcinomas." He went on to report that 18% of the 519 carcinomas diagnosed in his institute between 1974 and 1983 were pre-clinical cases detected by mammography, and over half of these cases were diagnosed on the basis of microcalcifications alone. Other studies have suggested that from 30% to 50% of breast cancers show microcalcifications on mammograms, while from 60% to 80% are found to have associated microcalcifications on histological examination (Murphy et al 1978, Sickles 1982). Typically microcalcifications appear as very small, sharp-edged blobs that are relatively bright in comparison with the surrounding normal tissue. They are generally between 0.01mm and 3.0mm in size and are of particular clinical significance when found in clusters of five or more in a 1cm x 1cm area (Sickles 1982).

Well-defined, or circumscribed, lesions – a second class of mammographic abnormalities – are considerably larger than individual calcifications and appear as large blobs with smooth edges that may be sharp or poorly defined. These

lesions can vary greatly in optical density on a mammogram with radiolucent (dark) cases generally indicating benign disease and denser radiopaque lesions often corresponding to malignancies (Tabar & Dean 1985). Circumscribed lesions may be observed in a variety of shapes, possibly appearing as smooth circles or ovals, or alternatively they may have a lobular appearance. In addition, the size of these lesions may vary from less than 1cm to greater than 10cm.

Stellate, or spiculated, lesions – a third type of mammographic abnormality – are often associated with malignancy. They generally appear as a distinct radiopaque mass fully or partially surrounded by radiating linear structures known as spicules. These spicules vary in length and may either radiate in all directions or appear bunched together like a "sheaf of wheat" (Tabar & Dean 1985). The central mass may contain circular or oval radiolucent patches, in which case the lesion is often benign. The ill–defined borders of stellate lesions may make them difficult to detect, especially when they are only a few millimetres in size, though these lesions can be several centimetres across.

Figure 1.2 illustrates some examples of abnormalities. Although these are the most common forms of mammographic abnormality, there are other indicators of breast disease such as architectural distortion, asymmetries between left and right breasts and thickening of the skin.

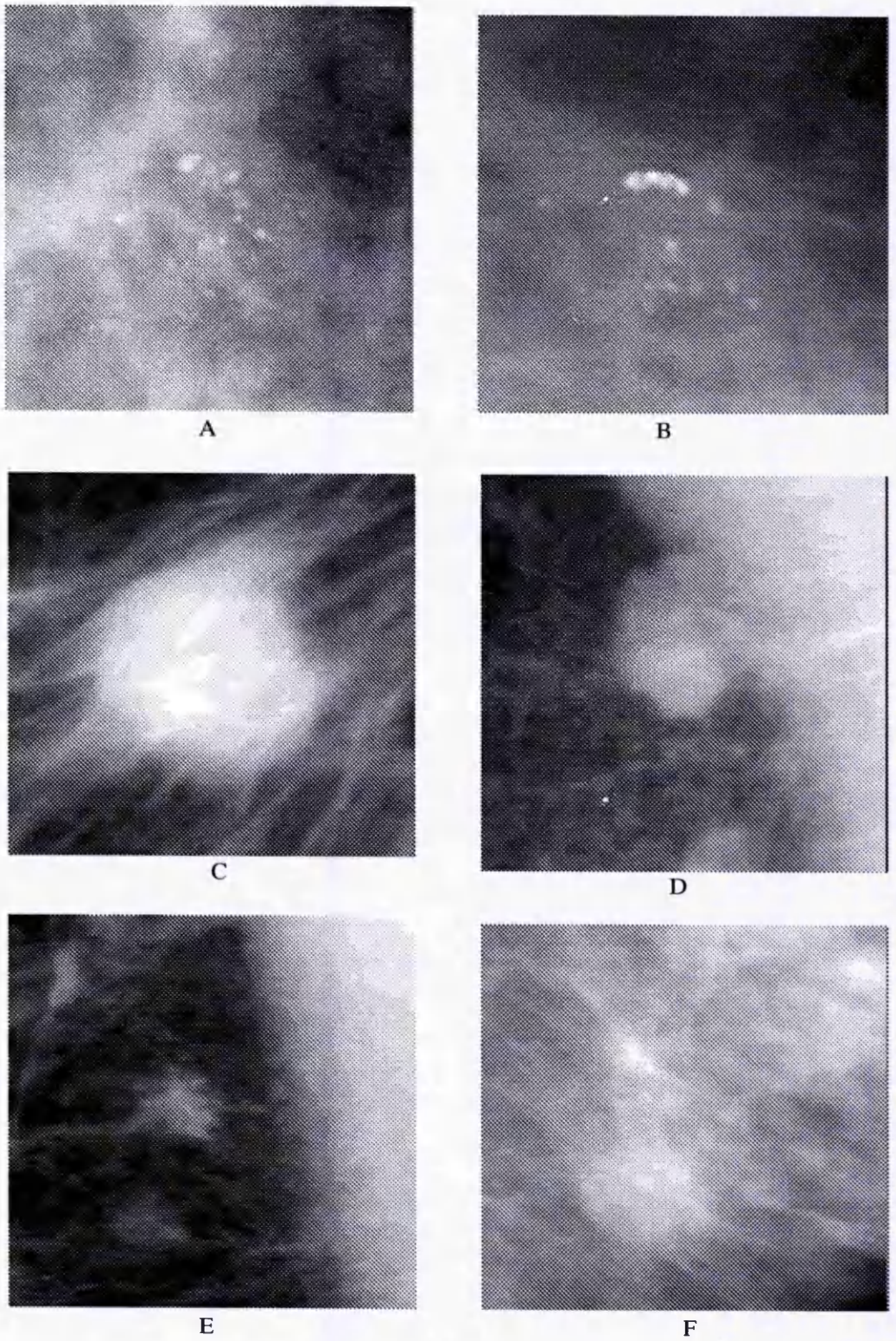

Figure 1.2: Examples of some Mammographic Abnormalities.
A and B are clustered microcalcifications, C and D are
well–defined lesions and E and F are spiculated lesions.

1.2.4 Possible Risks of Mammography

Although mammography has great value both as a diagnostic tool in symptomatic women and as a method for the detection of non–palpable carcinomas in apparently well women, some concern has been expressed about the exposure of the breast to X–ray radiation (Morgan 1985). Irradiation of the breast is of particular concern because the breast is one of the organs most susceptible to radiation carcinogenesis, even more so than bone marrow, the lungs or the thyroid gland (Gregg 1977, Bailar 1978).

Ever increasing sophistication in radiological techniques has led to the production of good mammographic images with very low doses of radiation (Lissner et al 1985), though the high susceptibility of the breast to radiation induced cancers means that a small risk may still exist. The actual radiation dose to the breast during a mammographic examination will depend both on aspects of the imaging system and characteristics of the breast, making it difficult to produce a single figure for the mean dose. However, Feig (1986) estimates the typical dose during mammography to be less than 0.1 rad, while Roebuck puts the estimate lower at around 0.05–0.015 rad.

Investigations of the risk associated with breast irradiation have studied populations who have been subjected to considerably higher radiation doses than are seen in mammography – doses of the order of 100–2000 rads (Feig 1986). Such groups have included the survivors of the atomic bombs at Hiroshima and Nagisaki (Tokunaga et al 1979) and women who have received radiotherapy for the treatment of benign breast disease (Baral et al 1977). These cases have shown definite increases in the risk of developing breast cancer at these very high doses, but the evidence for the risk at the low doses seen in mammography requires extrapolation from these results and is far less clear.

If there is a risk of developing breast cancer as a consequence of the low radiation dose received during mammography then it is so small that it has never been

observed (Roebuck 1986). Even if the estimated risk obtained from the models based on high dose studies were to be confirmed then it would be extremely low in comparison with other everyday activities – approximately equivalent to smoking one eighth of a cigarette or travelling ten miles by car (Feig 1986).

It is commonly held that the great benefits of screening by mammography outweigh the small potential risks involved (Morgan 1985), but in recognition of the possible hazard, a number of recommendations have been made concerning the age at which the screening of women in low risk groups should begin, particularly since higher doses are needed to image young, dense breasts. For example, Breslow (1977) suggests that mammography should not be performed routinely on women under the age of 50, while Lissner (1985) endorses clinical examination plus mammography for women over 30 years old. The Forrest report (1986) recommended mammographic screening using a single medio–lateral view of each breast for all women between the ages of 50 and 64 at three yearly intervals. It is the recommendations of this report that have been adopted in the UK by the National Health Service as the basis of a national screening programme.

1.3 Mass Screening

As has already been discussed (section 1.1) the effectiveness of the early treatment of breast cancer relies on the detection of the tumour at an early, pre–symptomatic stage. This necessitates the mass screening of apparently well women, which is now the current policy of the UK National Health Service.

One of the earliest, and most important, studies of the usefulness of mass screening for breast cancer was the Health Insurance Plan study (HIP), instituted in 1963. In this study, 62 000 women aged between 40 and 64 were randomly chosen and divided into two carefully matched groups, a control group and a study group, each consisting of 31 000 women. The women in the study group were

invited to attend an examination that employed both palpation and mammography. The two–thirds of the women in the study group who accepted the invitation were given an initial examination and three subsequent annual examinations.

In a follow–up study conducted nine years later it was found that 128 deaths from breast cancer had occurred in the control group, compared to 91 deaths in the study group (Shapiro et al 1973). This reduction in mortality rate of approximately one–third persisted after a twelve year follow–up.

Since the HIP study, there have been a number of large scale mass screening studies, all of which have, to a greater or lesser extent, demonstrated results in the form of reduced mortality rates among those women who have undergone regular screening. For example, Tabar and his colleagues (1985) have reported on a study begun in Sweden in 1977. In this case 162 981 women aged 40 years and over were randomly assigned either to a control group or to a study group that was offered screening by mammography every two or three years. At the end of 1984, there was a 31% reduction in the breast cancer mortality rate among the study group.

Results such as these demonstrate the effectiveness of mass screening for the detection of breast cancer at an early enough stage to allow significant improvements in the treatment of the disease.

1.4 Overview of Thesis

1.4.1 Objectives

The preceding sections have discussed the usefulness of screening for breast cancer, and the important role that mammography plays as a screening technique. Current NHS practice is to invite all women aged 50–64 for screening using mammography once every three years.

The effectiveness of mammography as a screening technique critically depends on the ability of a radiologist to detect the early signs of breast cancer in the mammographic image. This is not an easy task, as these abnormalities may be extremely subtle and they are embedded in the complex structured backgrounds associated with normal breast tissue.

Studies of the eye–movements of radiologists when they are reading radiographic images have revealed that the search patterns used are neither systematic nor complete (Kundel and Nodine 1978). It is often the case that regions of the image may not be fixated during the search and this can lead to abnormalities being missed. It is conceivable that this problem may be exaggerated during screening, when the radiologist is required to read a large quantity of films, typically 100 cases per session, the vast majority of which are free from abnormalities.

One possible application of computer vision to mammography is to develop algorithms that can automatically detect the early signs of breast cancer in digitised mammograms. These algorithms can then be used to generate attention cues, or 'prompts' that direct the attention of the radiologist towards suspicious regions of the image.

Several studies have suggested that prompting may be an effective method for improving the performance of radiologists in reading mammographic films, at least in experimental settings. However, there are a number of issues that require investigation before prompting could be considered an acceptable technique for use in a clinical environment.

One important question that should be considered is the way in which errors in prompt generation might affect the performance of the radiologist. It may be the case that, under certain circumstances, the radiologist could become overly confident of the accuracy of the prompting system. In such a case a false–positive prompt could possibly lead the radiologist to make a false–positive judgement

about the presence of an abnormality, when the film may have been correctly judged as normal had the prompt not been present.

Similarly, if a radiologist becomes too greatly reliant on the prompting system, the examination of the unprompted regions of the film may be less thorough than it might otherwise have been. This could have particularly serious consequences in the cases where abnormalities are missed by the prompting system. If the radiologist misses an abnormality that would have been detected if no prompting had been used, then prompting is clearly not a very effective technique.

The problem becomes compounded where the prompting system produces both false–positive and false–negative errors on a single film (or pair of films). In this case, the prompts will actually serve to direct the attention of the radiologist away from the location of the abnormality to some possibly clinically insignificant region of the film, which is precisely the opposite of the effect that prompting is designed to achieve.

Other problems may occur when the prompt generation system does not include algorithms for the detection of every type of mammographic abnormality. For example, a system may be set up to detect spiculated lesions and architectural distortions, but may not target any other forms of abnormality, in which case it is necessary to understand how the detection of untargeted abnormalities is affected by the prompting of the targeted ones.

Although it is possible that in the future a prompting system may be developed that can effectively target all of the possible classes of mammographic abnormality, it is extremely unlikely that such a system would be able to do so infallibly. Indeed, it should not be necessary for a prompting system to be infallible for it to be useful, since the radiologist could reasonably be expected to have a certain degree of tolerance to some small amount of error in the prompts. In order to develop an effective prompting system it is necessary to

understand how errors in the prompts may affect the detection performance of radiologists.

The aim of this study was to investigate some of the questions concerning prompting in mammography, particularly the effects of errors in prompt generation on the detection performance of radiologists.

1.4.2 Overview

The next chapter of this thesis will go on to examine the film reading process in more detail, looking particularly at the types of errors that radiologists make when searching for abnormalities and the cognitive mechanisms underlying visual search, especially the role of attention. In addition, some of the previous work on attention cueing and prompting will be discussed.

Chapter 3 introduces signal detection theory and ROC analysis – an important tool for measuring the detection performance of both human and artificial observers. The methods described in this chapter will be used throughout the rest of the thesis.

Chapter 4 will review some of the methods that have been developed for the automatic detection of abnormalities, particularly for the detection of clustered microcalcifications and tumour masses. Chapter 5 will then describe two particular detection algorithms in detail and compare their results when the algorithms are applied to a set of digital mammograms.

Chapter 6 describes an experiment carried out to investigate the effects that varying the accuracy level of the prompts had on their usefulness as aids to the radiologist.

The second experimental study, described in chapter 7, was designed to investigate the usefulness of prompting in a clinical setting. This was a study carried out at several screening centres around the UK.

Chapter 8 also describes an experimental study, this time using a simulated mammographic task. This experiment was designed to examine the relationship between the true–positive and false–positive prompting rates of successful prompting systems.

The fourth, and final, experiment described in chapter 9 was a large scale investigation carried out in a clinical setting. This study develops on the findings of the experiment described in chapter 8 and applies them to a realistic screening environment.

Finally, chapter 10 will draw together some of the ideas discussed in the thesis and attempt to draw some conclusions about prompting and the conditions required for the technique to be a useful aid to the radiologist.

Chapter 2

Mammographic Film Reading

2.1 Errors in Film Reading

Screening by mammography can clearly be of great value in the control of breast cancer. However, the effectiveness of mammography as a diagnostic tool relies on the ability of a radiologist to correctly interpret the mammogram. This is by no means an easy task since the signs of early breast cancer can be subtle and difficult to detect when embedded in a dense or complex structured background.

2.1.1 The Consequences of Error

Two general types of error are possible in the interpretation of mammograms; false–positives and false–negatives.

A false–negative error occurs when a mammogram containing an abnormality is classified as normal – the abnormality has been missed. In other words, a woman who is suffering from the early stages of breast cancer is diagnosed as healthy. This is clearly the most serious of errors, as the delay in correct diagnosis and consequent treatment may adversely affect the woman's chances of recovering from the disease.

Burns (1978) reported a false–negative rate of 7% among all patients with breast cancer who were examined by mammography at her institute. Among the false–negative group, correct diagnosis was delayed by between 4 and 260 weeks, with a mean delay of 45 weeks. A study by Thomas (1978) which compared various diagnostic techniques found a 19% false–negative rate when mammography was the sole diagnostic method used.

Lesnick (1977) studied the preoperative mammograms of 52 patients who had been diagnosed as having breast cancer, only two of whom had been diagnosed by mammography. When these mammograms were examined by radiologists, 29 of them were classified normal and 4 as having characteristics typical of benign tumours, yielding a total false–negative rate of a staggering 63%. It should be noted that the population under study in this case represented cases of breast cancer in symptomatic women, rather than the spectrum of disease detected by the screening of asymptomatic women.

The second type of error, false–positives, occur when a normal mammogram is incorrectly classified as containing an abnormality. Although this type of error is not potentially life–threatening, it does have important negative consequences, such as the psychological impact on the patient. Any positive diagnosis of cancer as a result of breast screening will fall upon a previously unsuspecting women who believed herself to be healthy.

The screening process, in the majority of cases, produces only low–level, transient anxiety. However, the levels of anxiety in women recalled for further assessment have been observed to be much higher (Hopwood & Maguire 1990).

Ellman and colleagues (1989) investigated the incidences of anxiety and depressive illness among a group of over 700 women who attended breast screening clinics, either for routine screening, for further investigation following a suspicious mammogram, or for investigation of breast symptoms.

They found that women in whom further screening showed no cancer (false–positives) had significantly higher anxiety levels at the time of the second screening prior to the establishment of a negative diagnosis when compared to those women undergoing routine screening in which no abnormality was found. However, in a three month follow–up study, they found that anxiety levels had fallen significantly in the false–positive group and were no longer any higher than those of the routine screening group. This suggests that though there is no lasting increase in psychological morbidity following a false–positive diagnosis that is subsequently found to be negative, there is a significant increase in anxiety before the negative result is established.

In Ellman's study, the false–positive diagnoses were found to be negative after further mammographic examination. This is not always the case. Lanyi (1985) reports that in 294 exploratory operations performed only on the basis of microcalcifications detected by mammography, only 50 carcinomas were found, which means that unnecessary surgery was performed in 83% of these cases.

Even if a false–positive diagnosis does not get as far as the operating theatre, the woman will still have to be recalled for a second mammogram. In addition to the waste of resources when a negative diagnosis is established, this will entail an additional dose of radiation to the breast.

2.1.2 The Causes of Errors

Since the consequences of errors in mammographic film reading can be severe it is worth trying to understand how and why such errors occur. One possible approach to this problem is to study the search behaviour of radiologists when they are reading films.

Kundel and Nodine (1978) studied the eye movement data of 5 radiologists who were scanning for small lung abnormalities in chest X–rays. This is a task that is,

to a certain degree, analogous to the detection of small abnormalities in mammograms since both tasks involve the detection of small subtle targets in complex backgrounds.

Kundel and Nodine found that the scanning patterns used by radiologists were neither systematic nor complete. In a large number of cases (around 40%) there was an initial circumferential sweep of the image with widely spaced fixations, a sequence that probably represents some global preliminary study. After this initial sweep, there was no consistently repeated sequence of fixations, although there appeared to be a certain amount of consistency in the locations that were fixated.

An interesting finding of this study was that the proportion of fixation time associated with given regions of the film image could be altered by giving verbal instructions before the task, usually in the form of a clinical history of the patient. Kundel and Nodine also found that there was a significant correlation between the proportion of fixation time spent on a region and the radiologists' subjective rating of the importance of that region.

These results led Kundel and Nodine to conclude that radiologists approach the film–reading task with an organised and highly selective search strategy that is biased towards those regions of the image that are considered to be the most informative. They suggest that pre–selection of potentially informative regions is based in experience and expertise and that the basic search strategy may be modified for a particular case when the radiologist is provided with specific clinical information on that patient. Using this model, Kundel and Nodine proposed three main sources of error in the interpretation of X–rays.

The first main class of errors were search errors. These occurred when the abnormality, a lung nodule in this case, did not fall within the scanpath of the reader. The scanpath was defined as the locus of the "useful field of view", which Kundel and Nodine believed to be an area of 4 degrees of visual angle centred on the foveal fixation point, though they do comment that 4 degrees is an average

figure, and that considerations of background complexity and task demands may alter the size of the "useful field".

The second group, recognition errors, occurred when the abnormality lay within the scanpath of the viewer but did not trigger recognition, in other words when the target was not disembedded from the background.

The final group consisted of decision making errors, which occurred when a suspicious region was located but was then misclassified.

Kundel and Nodine used dwell time as the criterion for distinguishing between recognition and decision making errors. They suggested that if the useful field of view fell on the abnormality for longer than 0.3 seconds then the target had been detected and consequently any error was of the decision making type, while if the dwell time was less than 0.3 seconds, then the abnormality had not been identified and a recognition error had occurred.

The data from a study of 20 errors and their classifications is shown in figure 2.1.

Error Type	Number	Proportion
Search	2	10%
Recognition	6	35%
Decision	9	53%

Figure 2.1: Classification of errors (from Kundel & Nodine 1978)

It should be noted that the sample size (n=20) in this study was limited and that a certain proportion of the total errors are unaccounted for. In addition it is not really possible to draw any definite conclusions about the errors in mammographic film reading from these data since the tasks involved are somewhat different.

Nevertheless, there are sufficient parallels between the two tasks to make the classification system a useful one, and it seems safe to conclude that a significant proportion of film reading errors may be due to inefficient search behaviour, or to oversights (Vernon 1971).

2.2 The Role of Attention

It is normally the case that an abnormality in a radiological image must be attended to in order to be recognised as such (Gale & Worthington 1986).

Usually, the locus of the observer's attention will correspond to the locus of foveal fixation, or more precisely, to a limited area around fixation often referred to as the "useful field of view" (Kundel & Nodine 1978). The size of the useful field would probably vary somewhat depending on the task, but a typical estimate is around 2.8 degrees of visual angle around fixation. Kundel and Nodine found that 90% of small lung abnormalities could be detected if they fell within this region.

Successful detection of an abnormality, therefore, seems to rely on directing the useful field of view to the location of the abnormality.

In the previous section it was mentioned that the radiologist approaches the film reading task with a pre–selected search strategy, or schema, which initially guides the pattern of fixations on the display. Once the search has been initiated, this schema does not remain rigidly fixed; it is subject to constant modification based on current foveal and peripheral visual information (Gale & Worthington 1986).

It may be the case that although foveal attention is required in order to identify abnormalities, peripheral attention plays an important role in guiding search through the image. This may be demonstrated by eye movement data in which large saccades from one part of the image to another are often observed. It seems as though peripheral attention is able to locate potentially suspicious regions within a relatively large area and direct the search of the image to bring high–acuity, foveal attention to bear on the region (Gale & Worthington 1986).

2.2.1 Feature Integration Theory

An interesting account of the role of attention in object perception is the feature integration theory proposed by Triesman (1985, 1988).

According to this theory, early or "pre-attentive" visual processing involves the parallel extraction of simple object properties such as shape, colour and orientation from the visual field. These simple properties are detected by hardwired "feature maps", each of which responds to a particular value of the feature dimension. For example, in the case of colour, separate feature maps might exist for red, yellow and blue. The parallel processing stage is then followed by a second, attentional stage involving the combination of the simple property information stored in the feature maps to form "objects".

Triesman and Schmidt (1982) suggest that the combination of features to form an object may be achieved by attending to the location of the object, so that attention acts to 'glue' the features together. Furthermore, they also suggest that features may be combined on the basis of stored knowledge, or in some cases randomly – producing "illusory conjunctions".

There are several lines of evidence that support the notion of feature integration theory (Humphreys & Bruce 1989), but perhaps the most interesting, and certainly the most relevant to mammographic screening, involve studies of visual search tasks.

Using feature integration theory as a model, Treisman (1985) makes a distinction between "feature" and "conjunction" targets. A feature target has one or more feature values that are not shared by any other elements in the display, for example, a red circle within a display comprising blue circles. A conjunction target, however, is defined by a combination of properties that are shared with other elements, such as a red circle within a display comprising a mixture of blue circles and red squares.

In a series of visual search tasks requiring a target present/absent judgement, Triesman found that the size of the display (number of distractors) had virtually no effect on performance when subjects were required to find a feature target. The target appeared to "pop–out" from the display, with detection being spatially parallel.

However, when a conjunction target was used, search latencies were seen to increase linearly with display size, suggesting that each item in the display had to be checked individually, or in other words, serial search was being employed, with each display element being attended to in turn.

Treisman suggests that in the case of feature targets, activity is produced in a separate feature map which is unaffected by the distractors, so that it is not necessary to check each individual element in the display, and "pop–out" occurs. With conjunction targets, there is no single feature map activated only by the target, making it necessary to attend to each item in the display in a serial fashion.

In another set of experiments, Treisman provides more direct evidence of the role of attention in object detection. The theory is that, if it is necessary to attend to a conjunction target in order to detect it, cueing the location of that target should direct attention towards it, allowing for detection without a serial search and consequently an increase in detection performance. However, cueing the location of a feature target should not significantly facilitate its detection, since feature targets are identified pre–attentively. The results of Treisman's experiments demonstrated this to be the case, with the brief presentation of valid pre–cue immediately prior to the display having a substantial benefit for the detection of conjunction targets and very little effect with feature targets.

An interesting aspect of feature integration theory is that it suggests a mechanism by which properties of the display may affect target detectability. In mammographic film–reading, abnormalities may appear with a variety of

different sizes, shapes and contrast levels. Similarly, the density and complexity of the background breast tissue may be subject to wide variations.

In some cases, clustered microcalcifications may be sharp-edged and fairly bright relative to the background tissue, and seem to have much in common with Treisman's feature targets, leading to rapid detection by pre-attentive, parallel search. However, in other cases, where the background is highly structured or the target/background contrast is low, clustered microcalcifications may behave more like conjunction targets, requiring systematic, serial processing of the image.

2.3 Prompting and Pre-cues.

It seems to be the case that attending to the location of a target is highly beneficial, if not essential, to the detection and identification of that target, and that in order to ensure that the target location is attended to, it may be necessary to conduct a systematic, serial search of the display. However, Kundel and Nodine (1978) have demonstrated that in a number of cases abnormalities are overlooked because they are not attended to. This suggests that the accuracy of radiological diagnosis might benefit if the attention of the radiologist is directed towards those regions of the image in which abnormalities are present. In other words, the radiologist may be "prompted" towards suspicious regions of the image.

The traditional procedure employed by experimental psychologists for directing attention towards a certain part of the display is the presentation of a brief pre-cue immediately prior to displaying the image. This procedure was developed by Posner (1978, 1980), who conducted a series of simple reaction time (RT) experiments in which subjects were required to press a button when a light appeared in any one of several boxes located at various points within the visual field. Prior to each target, the subjects were presented with one of two types of cue. In the control condition this was a cross in the centre of the display that gave no information concerning the target location, while in the experimental

condition the box in which the target would appear was briefly illuminated. The cue was valid on most trials, but on a small proportion it was invalid.

In order to separate the effects of prompting from the effects of eye movements, the cue was presented very briefly, so that the interval between cue onset and target onset was too short for eye movements to occur. The latency of eye movements is generally estimated to be about 180–200 msec (Posner 1980).

Posner found that that the RTs were faster than in the control condition when a valid cue was presented, but slower when an invalid cue was given, thus demonstrating the benefits of cueing target location in the detection of that target.

Several other researchers have found similar benefits of pre–cueing on the detection of targets (eg; Eriksen & Yeh 1985, Eriksen & Murphy 1987, Tsal 1983), which suggests that some form of prompting based on pre–cueing may be useful to the radiologist. However, before prompting in any form can be accepted as a useful aid to radiological screening, there are a number of issues which must be addressed.

Firstly, the great majority of the evidence supporting pre–cueing comes from studies that have been both highly artificial and relatively simple. In almost all cases both the target and the background, or distractors, have been clearly defined and responses have generally been either simple RT or discrimination tasks. The question arises as to how far the results of these studies can be extended to situations involving real and often highly complex images, such as those seen in mammography.

Secondly, in most attentional pre–cueing studies, measurement of performance has been in terms of reaction time. This is not really an appropriate criterion for improvements in mammographic screening, where performance benefits in terms of improved detection accuracy are required. Treisman's (1985) study did use accuracy, expressed in terms of the signal detection measure d' (see section 3.2.1)

as the measure of performance, and the pre-cueing of conjunction targets did lead to an improvement in detection accuracy. However, as mentioned above, the task in this case involved clearly defined artificial targets and distractors.

Furthermore, most studies of pre-cueing have involved the detection or location of only a single target in each display. In mammography it is quite possible for more than one abnormality to be present in an image. This potential for multiple targets requires an understanding of the effects of presenting a target with multiple prompts.

Perhaps the most important issue concerning prompting in mammography is the effect of presenting invalid prompts. If prompts are to be used to mark suspicious regions of a mammogram, then these areas of interest must be detected in some way, probably by some sort of computer vision system. Since no such system is likely to be completely specific, it follows that a certain number of false-positive, or invalid, prompts will be generated. Therefore it is important to know how these invalid prompts may affect detection performance.

Traditionally, it has been thought that the presentation of an invalid pre-cue has a detrimental effect on detection performance. This view has received a certain amount of support from experimental evidence, at least from simple RT tasks in which it is generally found that the presentation of an invalid pre-cue acts to increase reaction time (Posner 1980, Eriksen & Yeh 1985).

However, other experimental evidence has shown that this might not be the case. Treisman (1985), in her studies of the effects of pre-cueing targets, found that the presentation of an invalid pre-cue had little effect on the detection accuracy, relative to the no pre-cue control condition, of either feature or conjunction targets.

This result is encouraging since it suggests that invalid prompts might not adversely affect detection performance, but there is a problem with this

conclusion. The invalid pre–cues in Treisman's experiments were not, strictly speaking, false–positives, since even when an invalid cue was presented the target was still present somewhere in the display and consequently the subjects knew that they must continue to search for the target. Having responded (shifted attention) to the invalid prompt , the subject would then have to embark on a serial search for the target, just as if no pre–cue had been presented, which might explain the similarity between the results of the no cue and invalid cue conditions.

In these experiments the target and other display elements were clearly defined, so that a subject searching for a red circle who received an invalid pre–cue towards a blue square would be able to immediately disregard the cued element and embark on a serial search of the image. However, in mammography, the targets are not so clearly defined and the identification of an abnormality is often largely a question of interpretation. It has been suggested that many radiologists will disagree with their own previous interpretation of a mammogram one time in five (Gale et al 1979). It is conceivable therefore, that by focussing attention on a suspicious, but not abnormal, object in an image, an invalid prompt may lead the radiologist to make a false–positive judgement that would not have occurred if no prompt had been presented.

2.3.1 Prompting in Mammography

The first study to examine the effects of prompting specifically in mammography was conducted by Chan and her colleagues (1990). They used an automated detection system based on a difference–image technique and locally adaptive grey–level thresholding in order to identify microcalcification clusters in digital mammograms. This procedure, and other computer vision techniques used in the detection of microcalcifications will be discussed further in section 4.2.

Using the automated system, Chan achieved an accuracy level of 87% true–positive cluster detection with an average of four false–positives per image,

which was referred to as "level 1 accuracy". She then simulated an accuracy level of 87% true-positives with 0.5 false-positives per image, which was described as "level 2 accuracy".

A set of 60 mammograms, half of which contained clustered microcalcifications, were processed as described above and printed out on a laser printer. Each image was printed out 3 times; once in a digitised but unprocessed form and once processed at each level of accuracy. The results of the computer detection system were superimposed on the processed films in the form of small open circles corresponding to the locations of the identified clusters.

The processed mammograms were then presented to 15 expert subjects who were asked to determine the presence and location of any microcalcification clusters in each image.

Using ROC analysis (see section 3.2.2), Chan found that both levels of prompting accuracy led to significantly higher detection performance than the control version without prompts, and there was no significant difference between the two prompting conditions.

These results imply that the presentation of valid cues to the locations of the abnormalities can significantly improve detection accuracy regardless of the number of invalid cues that are also presented. Although this is an encouraging result, it should be noted that there are certain questionable aspects of the methodology.

Firstly, the study time per image was limited to five seconds, a figure that was calculated from the suggestion that a full four-image mammographic study can be screened in as little as 45 seconds (Sickles et al 1986). A limited study time of this nature does not reflect the real screening task, in which the only time limits faced by the radiologist are self-imposed.

Secondly, the mammograms were only presented in the digitised form as hardcopies. It is quite possible that the loss of spatial resolution inherent in the digitisation process may adversely affect the detectability of subtle microcalcifications, which may have affected the results. Certainly this casts some doubt on the applicability of these results to prompting with conventional mammographic films.

In addition, perhaps the most significant problem with Chan's design was that each condition was presented to the subject as a separate block of 20 images, and prior to each block the subject was informed of the condition to be used and the level of accuracy involved. As has already been discussed, Kundel and Nodine (1978) found that the scanning strategy of a radiologist is affected by verbal instructions concerning the information to be presented and the nature of the task. It is possible therefore, that in Chan's experiments the search biases of a given subject differed depending on the experimental condition, and that since the task was slightly different in each condition, slightly different scanning strategies were used in each case. For example, it might be the case that the subjective ratings of confidence that a cluster was present may have taken a different interpretation for each block.

Nevertheless, the results do indicate that a computer detection system can be used to produce a statistically significant improvement in the accuracy of mammographic film reading, a goal that is certainly worth pursuing.

Chapter 3

Experimental Methodology

The aim of this section is to give a general introduction to the topics of signal detection theory and receiver operating characteristic (ROC) analysis, which form the basis of the methods used for testing the performance of both radiologists and automatic detection systems.

3.1 Introduction to Signal Detection Theory

3.1.1 Classical Psychophysics

It is generally the case that changing the value of a stimulus will produce a change in the perception of that stimulus. For example, an increase in the intensity of light will lead to an increase in perceived brightness. The relationships between stimuli and sensations are the province of psychophysics, a term first used by Gustav Fechner (1860).

Fechner was the first to adopt a mathematical approach to describe the relationship between mental experience and the physical world. The basic model of this relationship is the psychophysical function, the nature and measurement of which is the concern of one branch of psychophysical study.

The other branch of psychophysics is concerned with the measurement of sensory thresholds. Fechner developed methods for estimating the just noticeable difference (*jnd*). Also known as the difference threshold, the *jnd* is the minimal difference between two stimuli that can be perceived as a change in sensation. The underlying assumption of this work was that the *jnd* is a fixed value that represents a fundamental unit of experience.

Unfortunately, there were two major problems with the measurement of the *jnd*. Firstly, the threshold did not appear to have a single value, but rather a range of values over which the probability of detection moved from 0% to 100% (Lloyd 1984). Secondly, values of the *jnd* varied depending on the method used to measure them.

The problems associated with threshold measurement led to the development of an alternative approach, signal detection theory, which rejects the concept of thresholds completely.

3.1.2 Signal Detection Theory

Signal detection theory, originally formulated by Green and Swets (1966) is based on the idea that any signal is presented against a background of noise that varies randomly about some mean value. When a stimulus is presented, the activity that it creates in the sensory system is added to the noise existing at that moment. This noise may be within the system itself or it may be part of the input pattern. The task of the observer is to determine whether the level of activity in the system is due to noise alone or the result of a stimulus added to the noise.

Figure 3.1 shows the underlying distributions of the signal detection task. M_1 and M_2 are the means of the "noise alone" and "signal + noise" distributions respectively. The activity level marked c represents the criterion level of the observer, ie. the level of activity that must be exceeded in order for a "signal present" decision. The criterion level is discussed further in the section 3.1.4.

In figure 3.1 the separation of the distributions, given by M_2–M_1, is a measure of the sensitivity of the system (MacMillan 1991). If the signal is strong (a high signal–to–noise ratio), then the separation will be large and the sensitivity of the system to that signal will be high. However, a weaker stimulus will lead to a greater probability that activity in the system could result from noise alone and sensitivity to the signal will be low.

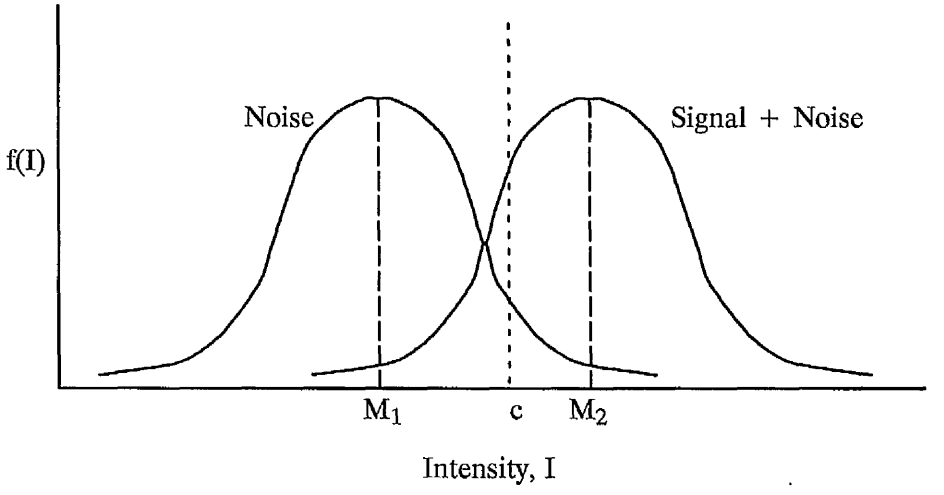

Figure 3.1: Underlying distributions in signal detection task

The detection of a signal is therefore a statistical matter and the observed gradual change from 0% to 100% observed in threshold measurement is just what would be expected (Lloyd et al 1984).

In addition, by separating the behaviour of stimuli from the decision–making process, signal detection theory allows useful measures of performance that are independent of the procedures used to calculate them. From figure 3.1 it is clear that the sensitivity of the system (M_2–M_1) is independent of the criterion level adopted by the observer. This means that while different procedures or differences in motivation may affect the criterion level of the observe, the

sensitivity of that observer to a given stimulus should remain fixed (MacMillan 1991).

The detection of an abnormality in a radiological image can be considered to be the problem of detecting a signal embedded in a background of structured noise – in this case normal breast tissue. This suggests that signal detection theory may be an appropriate paradigm for studying the performance of both human observers and artificial systems engaged in the film reading task.

3.1.3 Measuring Sensitivity

The signal detection theory approach covers a range of experimental protocols, the simplest of which is a forced discrimination task. In a typical example of this type of study, an observer might be presented with a series of stimuli each consisting of either noise alone or noise plus a target signal. At each presentation the observer might have to respond either "signal present" or "signal absent". Responses can thus be classified as one of four types as indicated in figure 3.2.

		Response: (Signal present ?)	
		"Yes"	"No"
Stimulus:	Present	HIT *(True–positive)*	MISS *(False–negative)*
	Absent	FALSE ALARM *(False–positive)*	CORRECT REJECTION *(True–negative)*

Figure 3.2: Possible responses in a simple signal detection theory experiment

As can be seen from figure 3.2 there are two possible types of error associated with this task. The first is a 'miss' which is when the observer fails to detect the

signal when it is present. The second is the 'false alarm' which occurs when the observer responds "signal present" even though there is no signal.

The performance of the observer can be summarised in terms of the "hit rate" – the probability of correctly responding "signal present" and the "false alarm rate" – the probability of responding "signal present" when there is no signal. From these measures it is possible to calculate the detection sensitivity of the observer, generally expressed in terms of the sensitivity index; d', which is calculated using equation 3.1.

$$d' = z(H) - z(F) \qquad [3.1]$$

where H and F are the hit rate and false alarm rate respectively, and z( ) is the inverse of the normal distribution function.

The z transformation converts the hit and false–alarm rates into units of standard deviation in such a way that a proportion of 0.5 has a z–score of 0, larger proportions have a positive z–score and smaller proportions have a negative score.

If an observer shows no discrimination at all, H = F and d' = 0. In this case the observer is operating at chance level and any positive judgement is as likely to be true as false.

Perfect accuracy, however, implies an infinite value of d'. In order to avoid this problem it is common practice to convert true–positive and false–positive rates of 0 and 1 to 1/2N and 1–1/(2N) respectively, where N is the number of trials (MacMillan 1991). For example, 25 hits and 0 misses gives H = 1 and F = 0. These values would be converted to H = 0.98 and F = 0.02, implying 24.5 hits and 0.5 misses.

3.1.4 Measuring Response Bias

Signal detection theory assumes that an observer has a fixed sensitivity for any given discrimination task. However, the willingness of that observer to respond "yes" rather than "no" may alter under different experimental conditions. For example, if the costs to the observer are greater for one type of error than for the other, then the observer may bias his/her response to reduce the possibility of making the more costly error. This is known as *response bias* and it reflects the strictness of the criterion level that the subject is using to determine an appropriate response.

The bias measure used in signal detection theory is c (for "criterion") and is calculated using equation 3.2.

$$c = -0.5[z(H)+z(F)] \qquad [3.2]$$

Where H and F are the hit and false-alarm rates and z() is the inverse of the normal distribution function. It should be noted that a negative multiplier is included in equation 3.2. By convention a positive bias is a tendency to say "no signal".

3.1.5 Signal Detection Theory Methodologies

It is a theoretical assumption of signal detection theory that, for a given observer on a given task, sensitivity will remain fixed even though response bias may change. In theory then, it should be possible to take a measure of sensitivity at a given criterion level and assume that it applies at all other criterion levels. However, it is often the case that in an experimental task, sensitivity may be subject to some variation at different levels of response bias. In practice it is often more useful to look at performance over a range of response bias levels and calculate an overall measure of sensitivity. This allows the sensitivity of an observer to be represented graphically by means of a receiver operating

characteristic (ROC) curve. ROC curves are discussed in more detail in section 3.2

One way of measuring different levels of response bias is to present the same stimuli to the observer in several trials, and request response criteria of varying strictness, for example; "Answer yes if a signal is definitely present", "Answer yes if a signal is probably present", "Answer yes if a signal is possibly present", etc. This method has two main drawbacks when used with human observers. Firstly, it is difficult to retain consistency in the observer's definition of the response criteria. Even in a single trial the observer's definition of precisely what constitutes "probably present" may differ from one stimulus to another. Secondly, this method requires that the observer undertake a great many presentations, possibly leading to a decrease in vigilance, and that the stimuli are repeated, so learning may interfere with the results. However, neither of these problems apply when this method is used to assess the detection performance of an artificial system and consequently this paradigm is often used for such purposes.

An alternative to the repeated presentations method is to use a rating scale. This is often a more effective method for assessing the performance of human observers. Rather than a simple Yes/No judgement, the observer is asked to rate his/her confidence that the signal is present on a scale that might range from "definitely present" to "definitely not present" with a number of points in between. Using this method, a single series of presentations can be used to generate measures of several levels of response bias, as indicated by the various scale points.

3.2 Receiver Operating Characteristic Analysis

3.2.1 Properties of ROC curves

According to signal detection theory, on any given discrimination task, an observer should have a fixed sensitivity but may vary in response bias. The locus

of true–positive rate and false–positive rate pairs that yield a constant value of d' is variously called an isosensitivity curve (Luce 1963), a relative operating characteristic (Swets 1973) or by engineering nomenclature, a receiver operating characteristic (ROC). This latter term will be used for the remainder of this text.

Figure 3.3 shows a set of ROC curves representing different levels of sensitivity. Each curve connects points with a constant value of d'.

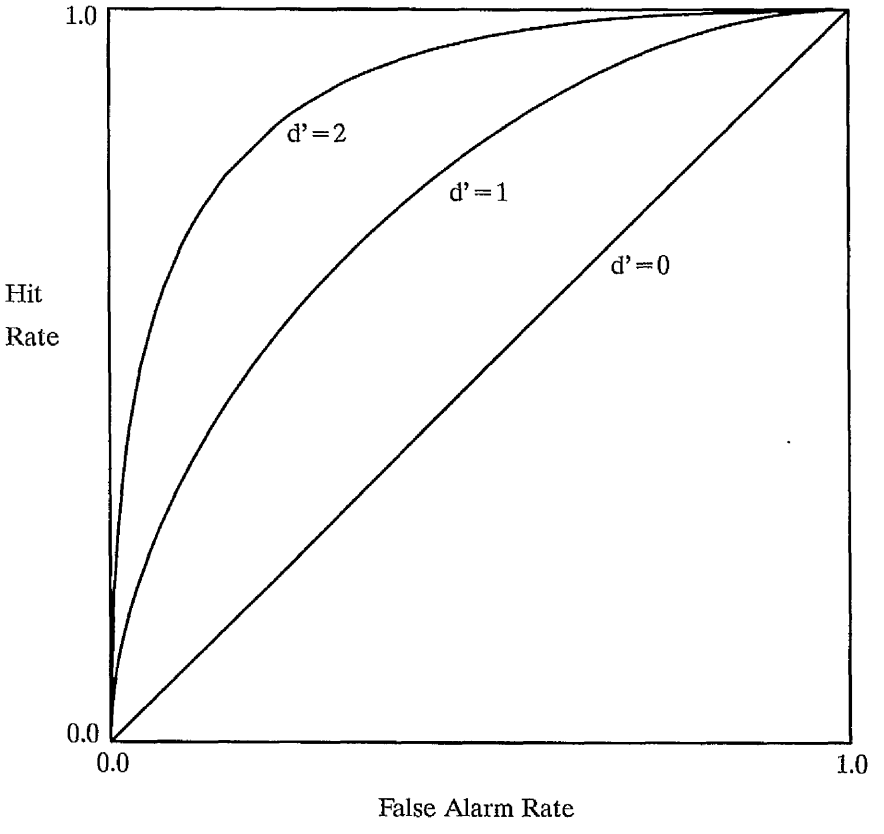

Figure 3.3: Set of ROC curves showing different levels of sensitivity

In figure 3.3, the curve marked d'=0 is the chance line at which the observer is showing no discrimination. Sensitivity increases as the curves move closer to the upper–left corner.

An important characteristic of the theoretical ROCs shown in figure 3.3 is that complete success in recognising one stimulus class is achieved at the cost of complete failure in recognising the other, ie. a hit rate of 1.0 can only be achieved with a false–alarm rate of 1.0, while a false–alarm rate of 0.0 can only be achieved with a hit rate of 0.0. ROCs that pass through (0,0) and (1,1) in this way are called regular curves (Swets & Pickett 1982). ROCs derived from experimental data, or empirical ROCs, are not always regular.

An alternative way to view ROC curves is as straight lines obtained by transforming the axes into z–scores. Figure 3.4 shows a set of such straight line ROCs plotted on transformed axes. On a straight line ROC the value of d' can be read directly as the intercept of the line on the y–axis.

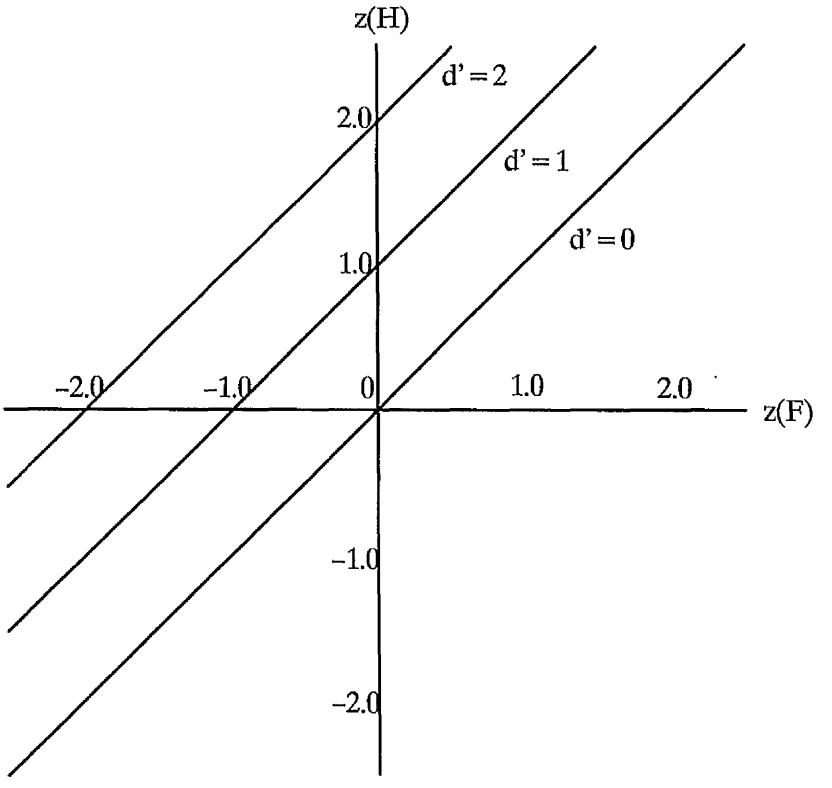

Figure 3.4: Straight line ROCs plotted on transformed axes

The theoretical straight line ROCs shown in figure 3.4 all have a gradient of 1.0. This means that the magnitude of d' corresponds to both the x–axis intercept, which may be referred to as d'_1, and the y–axis intercept, d'_2. The ROC has unit slope because the underlying 'noise' and 'signal+noise' distributions, as illustrated in figure 3.1, have equal variance. Again, this is not always the case with empirical ROCs.

3.2.2 Empirical ROC analysis

The curves shown in figure 3.3 and 3.4 are, of course, theoretical curves and it is likely that in many cases experimental data will not fit into such a straightforward pattern, with each level of response bias (each point on the ROC curve) producing an equal value for sensitivity. However, sensitivity indices are useful for comparing systems since they can be used in statistical tests such as t–tests and analysis of variance (ANOVA) in order to determine statistical significance. Therefore it is useful to be able to calculate such measures from experimental data.

The first step in calculating sensitivity measures from experimental data is to generate a straight line ROC by converting the values for false alarm and hit rates into z–scores and then plotting them on linear axes. The straight line that passes through these points, which may be calculated by regression analysis, is a straight line ROC.

With the theoretical straight line ROC curves in figure 3.4, the value of d' could be read off from either the x–axis intercept (d'_1) or the y–axis intercept (d'_2). However, this is only appropriate when the underlying "noise" and "signal + noise" distributions have equal variance, and the resulting straight line ROC has a slope of 1.0. If the slope of the ROC is other than 1.0 then an alternative measure of sensitivity must be calculated.

Figure 3.5 shows an example of a straight line ROC that might be derived from experimental data. The line has a gradient other than 1.0 so different results are given by d'_1 and d'_2. Both of these values are inaccurate as d'_1 overestimates the sensitivity of the system and d'_2 underestimates it. An alternative measure, D_{YN}, has been suggested by Schulman and Mitchell (1966). D_{YN} is simply the perpendicular distance between the ROC line and the origin and is easily calculated, but unfortunately gives values that are always smaller than both d'_1 and d'_2.

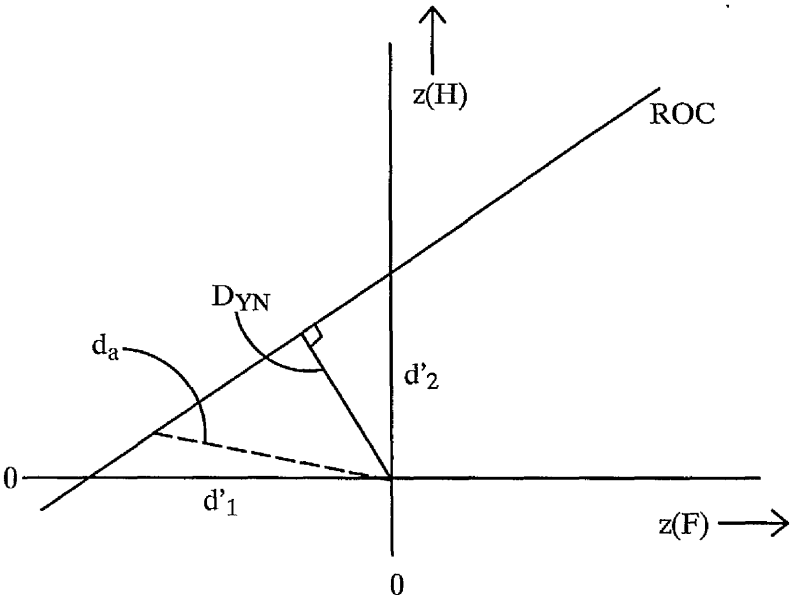

Figure 3.5: Sensitivity measures for non-unit slope ROC

A more appropriate measure, d_a, and was originally developed by Simpson and Fitter (1973). This is calculated by multiplying the perpendicular distance, D_{YN} by $\sqrt{2}$ and represents the hypotenuse of an equilateral triangle whose other two sides have length D_{YN}. In addition to having a value that is intermediate in size between d'_1 and d'_2, $d_a = d'$ when the ROC has unit slope. The index d_a may be calculated from equation 3.3.

$$d_a = [\, 2 / (1+s^2) \,]^{1/2}\, d'_2 \qquad [3.3]$$

where s is the slope of the straight line ROC and d'_2 is the y–axis intercept.

Another useful measure of sensitivity, A_z, can be calculated using d_a (or D_{YN}). A_z is an estimate of the area under an ROC curve. For a standard ROC curve, sensitivity is higher the closer the curve comes to the upper–left corner of the ROC space. This implies that the greater the area under the curve, the higher the sensitivity of the observer. A_z is expressed as a proportion of the total area of the ROC space, so that $A_z = 0.5$ represents chance level and $A_z = 1.0$ represents perfect performance. The index A_z can be calculated from equation 3.4.

$$A_z = \Phi(d_a / \sqrt{2}) \qquad [3.3]$$

where $\Phi( )$ is the normal distribution function, the inverse of the z–transformation used to calculate the sensitivity measure d' and the criterion level c (see equations 3.1 and 3.2).

3.2.3 ROC curve variants

The preceding sections have discussed the application of ROC analysis to experimental data. Although the methods described may be applied to a wide range of discrimination, detection and classification tasks, in order for ROC analysis to be used it is generally the case that the observer be restricted to a single response for each presentation.

In some cases, it is inappropriate to impose such a restriction on the observer. For example, a computer–based detection algorithm may be attempting to locate a target in a digital image. The algorithm may flag several locations in the image, one of which is the target while the others are false–positives. In terms of ROC analysis as discussed in the preceding sections, this would be considered a true–positive as the target has been detected. However, this ignores the data concerning the false–positive locations that were also flagged. In order to take this

false–positive data in to account an alternative to the standard ROC analysis is required. This alternative is known as the the free–response operating characteristic (FROC) and is illustrated in figure 3.6.

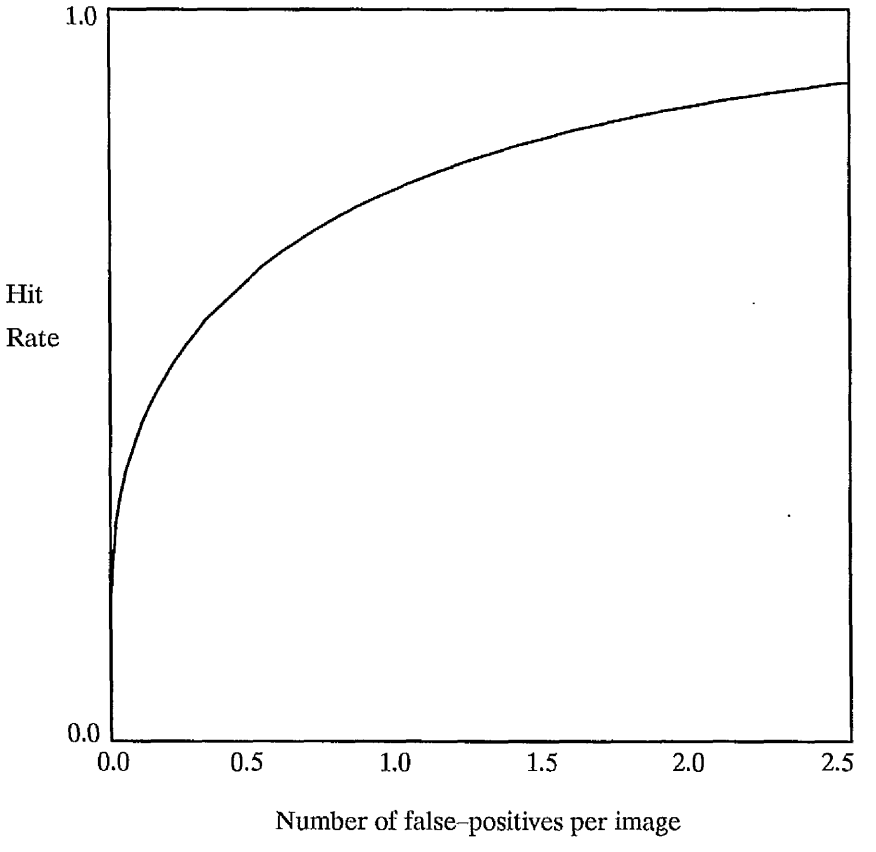

Figure 3.6: Example of free–response operating characteristic (FROC)

For an FROC the hit rate is still expressed as a proportion, as it is still appropriate to express the number of hits in this way. However, by accounting for all possible false–positives, it becomes inappropriate to express the false–positives as a proportion since there is theoretically no limit to the number of false positives that may be generated. Therefore, the false–positive rate is expressed as the number per image.

Note that in figure 3.6 the FROC curve does not pass through the upper–right corner where the the hit rate would be 100%. This is often the case with FROC curves where the number of false–positives generated in order achieve 100% true–positive detection would be extremely large.

In some cases FROC curves provide a clearer picture of how a system is operating than a conventional ROC curve. However, no very useful measures of sensitivity can be derived from FROC curves. They are really just a tool to illustrate a system's performance in isolation. If it is necessary to derive some statistical measure of performance, for example if two systems are to be compared, then the FROC must be converted to a conventional ROC curve.

Converting FROCs to ROCs is easily achieved by re–examining the original data and assigning a single response class to each stimulus presentation. In other words, if any location is flagged on an image, then this is considered a positive decision. The problem with this conversion is clear. If the algorithm in the above example is presented with an image containing a target and it flags a location that is not the target then this will be considered a true–positive response, as the system has responded positively to a target–bearing image.

In order to avoid this problem another slight variant on the ROC is required. An 'ROC with localisation' (LROC) appears exactly the same as a conventional ROC and is generated when an FROC is converted into the conventional form. To generate an LROC the conversion is similar to that described above with the restriction that a response can only be classed as true–positive if a genuine true–positive is present, ie. the correct location as well as the correct image has been flagged.

The LROC generally gives a lower estimate of sensitivity than the ROC and is most useful for extracting statistics for comparison with other systems whose

performance has been analysed in a similar way. The best illustration of performance in these cases is still the FROC.

Chapter 4

Computer Vision in Mammography

This chapter is intended to introduce some of the problems associated with digital mammography and provide a review of some of the methods that have been suggested for the automatic detection of two important abnormalities observed in mammograms; clustered microcalcifications and tumour masses.

4.1 Digital Mammography

4.1.1 Issues in Digital Mammography

Digital mammograms may be acquired by one of two methods; primary and secondary digital mammography. Primary digital mammography involves recording the emitted X-ray beam directly into a digital format, completely by-passing the need for a conventional film mammogram, while secondary digital mammography involves taking a conventional mammogram and digitising it with an appropriate scanner or CCD camera. Although the technology is in place in some institutions for the acquisition of primary digital mammograms, the use of such techniques is far from widespread and the vast majority of research in this field uses secondary techniques. All of the images discussed in this text have been acquired by means of secondary digital mammography.

The digitisation of mammographic images offers a number of potential advantages when compared to standard screen–film procedures. Digitised images may readily be stored within a computer system, allowing for improvements in data security and display flexibility and facilitating the rapid transmission of the digital images between clinicians at different stations. Additionally, image processing techniques may be applied to the digital mammogram in order to enhance the image or allow automated film reading. However, these advantages must be offset against a significant problem: the loss of spatial resolution that is inherent in the digitisation process.

This reduction in spatial resolution can lead to a reduction in target detectability and consequently a reduction in the detection performance of the radiologist (Chan et al 1987a). Small abnormalities, such as microcalcifications, may fail to be adequately represented if the sampling frequency is insufficient. Similarly, the boundaries between adjacent microcalcifications may be blurred by the digitisation process, so that two or more microcalcifications may appear as a single structure.

The problems associated with the digitisation of mammographic images increase as the spatial resolution decreases. Fam and her colleagues (1988) applied a computer detection algorithm to images digitised at sampling rates of either 0.1mm per pixel or 0.2mm per pixel. Fam found that the detection performance of the algorithm was better at the higher sampling rate. However, even at relatively high resolution the detection performance of the radiologist is still observed to be better with the original film than with the digitised image (Chan et al 1987a).

A more recent study by Karssemeijer and his colleagues (1993) presented 10 radiologists with 72 image patches containing microcalcifications and asked them to classify the patches as benign or malignant and as comedo or non–comedo. These were presented both as the original film and in a digitised form with a

resolution of 0.1 mm per pixel. There appeared to be no significant difference between the performance of the radiologists using the original or digitised versions, though it should be noted that this was a classification task and did not involve detection.

Nab (1992) presented radiologists with 210 mammograms; 135 normal, 75 containing lesions and 60 containing microcalcifications. All were digitised at 0.1mm per pixel and presented in both the original and digitised forms. Again, there was no significant decrease in performance with the digital images, and in this case there was a detection task involved. It appears therefore that images digitised with a sufficiently high resolution do not adversely effect the performance of the viewing radiologist.

Research in the field of computer vision has led to the development of increasingly sophisticated methods of image processing. These procedures can readily be applied to digital mammograms in order to enhance the images and facilitate the detection of abnormalities. Alternatively, systems may be developed for the automated detection of abnormalities. A fully automated screening system is still a rather distant goal, since it is very difficult to determine automatically that a film is unequivocally normal – i.e. free of every type of mammographic abnormality. However, the results of a computer based detection system could be used to assist radiologists in the screening task. Ideally, computer–aided diagnostic (CAD) systems could be used to assist the radiologist to such an extent that detection performance is at least as good, if not better than that observed with the original mammograms.

The remainder of this chapter will discuss the ways in which image processing techniques have been applied to the problem of detecting abnormalities in mammograms. Firstly, algorithms for the detection and enhancement of microcalcifications will be reviewed. The next section will then examine some of the methods that have been used to detect lesions. Two particular systems have

been implemented as part of this study, one for the detection of microcalcifications that combines evidence from a number of cue generators and the other which is based on gaussian pyramids and is designed to detect both microcalcifications and lesions. The final section will describe these two systems and present the results obtained from testing their detection performance.

4.1.2 Image Enhancement

The aim of image enhancement is to facilitate the detection of abnormalities in an image by highlighting those features of clinical significance while suppressing parts of the image that have no diagnostic value, such as normal background tissue.

One such procedure was employed in a study conducted by Chan and her colleagues (1987a). Chan used a set of 32 images, 12 of which contained subtle microcalcifications. These images were digitised with a sampling rate of 0.1mm per pixel and enhanced using an unsharp mask filter. Unsharp masking is a simple image processing technique that selectively enhances a certain range of spatial frequencies. The range to be enhanced is specified by the size of the mask, which in this case was 91x91 pixels. Enhancement of the high spatial frequency components of the image should have led to an increase in the contrast of small structures such as microcalcifications.

Each image was presented in each of the three forms (original, digitised–unprocessed and digitised–unsharp masked) to a group of nine radiologists. Chan found that the enhancement using the unsharp masked filter resulted in an improvement in the detection of clustered microcalcifications when compared to the unprocessed digital images, though the level of improvement was only barely statistically significant. However, neither the enhanced nor the unprocessed digital images led to a detection performance as high as that observed with the original images. It seems, therefore, as though this particular

enhancement procedure, while effective to some degree, was not effective enough to counter the problems associated with the digitisation process.

Dhawan and his team (1986) investigated a range of contrast enhancement procedures based on optimal adaptive neighbourhood processing. The first stage of this system involved establishing an optimal neighbourhood around each pixel. A neighbourhood consisted of a central kernel of adjacent pixels and a single pixel annulus a short distance outside this kernel. The size of the neighbourhood was determined by finding the size at which a local contrast function reached a maximum.

The transformed value of each pixel was based on a number of factors, including the original pixel value, the average density of the annulus and the result of applying a contrast enhancement function to the local contrast value. Dhawan examined several methods of contrast enhancement, including the square root, exponential, logarithmic and trigonometric functions.

The effectiveness of each of the enhancement functions was measured using histogram analysis to examine the resulting grey level distribution. This analysis led to the conclusion that logarithmic and exponential functions were most effective for contrast enhancement, while the square root function enhanced the noise, resulting in the degradation of the image.

Unfortunately, the results of the enhancement processes were not presented to any radiologists, so it is not possible to determine whether or not the enhanced parts actually corresponded to clinically significant features in the image. Furthermore, Dhawan reports that the computing time required for this type of processing was substantial.

However, the notion of adaptive neighbourhood processing has a certain amount of appeal. The complexity and variability observed in mammographic images suggest that this sort of locally adaptive method may be more effective for image

enhancement than global processing such as the unsharp mask filtering employed by Chan.

As was mentioned earlier, image enhancement is not the only possible application of CAD. An alternative application involves the detection of abnormalities. The remainder of this section will examine some of the techniques that have been developed for the detection of microcalcifications in digital mammograms.

4.2 The Detection of Microcalcifications

4.2.1 Pattern Recognition

One of the earliest microcalcification detection programs was developed by Wee and colleagues (1975) for the purposes of classifying benign and malignant lesions. The program employed a pattern recognition algorithm that operated on feature values extracted from microcalcifications that had been identified by edge detection and subsequent boundary tracing. Edge detection was achieved by adaptive local thresholding, with thresholds being established from the average grey level in the region under analysis.

The pattern recognition system used a set of seven features to classify the microcalcifications. These features included area, average grey level, contrast and smoothness, and using all seven features the system exhibited a classification accuracy of 88.2% with a set of 51 images containing microcalcifications (28 benign, 21 malignant). Wee also found that an accuracy level of 84.3% could be attained using just three features; average grey level, horizontal length and contrast.

It should be noted that there are certain questionable aspects of this study. Firstly, Wee mentioned using some form of preprocessing methods to sharpen and smooth the image prior to edge detection, though no details of these methods

were given. More significantly, the same set of images was used for both training and testing the system, which raises some doubt as to the validity of the results.

A rather more sophisticated pattern recognition system was developed by Davies and Dance (1990) for the automatic classification of normal and abnormal mammograms. The overall structure of this system was similar to that developed by Wee; local adaptive thresholding followed by feature extraction and pattern recognition. However there are a number of important differences between these two programs.

The first step in Davies and Dance's system involved pre–processing with a mode filter. This filter set the value of a pixel to zero if the modal value of its adjacent pixels was zero, and should have served to filter out the background area of the digital mammogram that did not conform to breast tissue.

The adaptive local thresholding procedure operated on regions that overlapped in such a way that any given pixel appeared in five different regions. The threshold of each region was determined from the grey level histogram of that region and a pixel was accepted in the final segmented image if it was still present in a pre–determined number of regions after thresholding. Davies and Dance refer to this pre–determined number as the "threshold overlap number".

Initially the image was processed with a threshold overlap number of 3. The structures in the segmented image were then analysed and five features were extracted; area, mean grey level, two shape parameters and edge strength. These features enabled the pattern recognition system to discriminate between microcalcifications and other image features. A clustering principle was then used to locate groups of three or more microcalcifications with nearest neighbour distances of less than 5mm.

The next stage involved segmenting the original image again with a threshold overlap number of one. Feature extraction was then used to establish the parts

of the image that corresponded with normal breast structures, such as ducts and blood vessels. The two processed images were then compared and those clusters in the first image that corresponded to normal breast tissue in the second were rejected.

Davies and Dance trained their system on a set of 25 images, then tested it with 50 different images, half of which were normal and half of which contained clusters. The results of their study were encouraging. They reported that their system successfully detected 47 out of 49 clusters, giving a true–positive rate of approximately 96%. The system also found a total of 9 false–positive clusters in the 50 images, which is less than 0.2 false–positives per image. In terms of film classification, the system correctly classified all of the abnormal images and 92% of the normal images. This result may have been due to the substantial weighting against false–negative decisions (missed clusters) that was built into the cost matrix of the pattern recogniser.

It seems, therefore, as though the rather more sophisticated system of Davies and Dance performed more accurately than that developed by Wee. It should be stressed, however, that the two systems were designed for somewhat different tasks. It may be the case that distinguishing between benign and malignant clusters is more difficult than deciding whether or not a cluster of any sort is present. These two tasks, detection and interpretation, represent the two main problems associated with mammographic film reading (Astley et al 1992), and the two systems demonstrate the way in which computer vision techniques may be applied to either of these tasks.

4.2.2 Feature Testing

In order to distinguish between clinically significant structures and normal breast structures it is necessary to examine the features of those structures. The pattern recognition systems described in the preceding section used feature analysis to

train the system to recognise the characteristic feature values of microcalcifications, allowing subsequent testing by comparison with these values.

The systems that will be examined in this section use feature analysis in a slightly different way. These systems are set up with pre–determined feature values or ranges of values that are accepted as characteristic of microcalcifications. These values form the basis of a set of tests, all of which must be passed in order for a cluster to be accepted. This approach is clearly illustrated in the work of Fam and her colleagues (Fam et al 1988, Fam & Olsen 1988).

The first set of three tests in Fam's system was applied to each pixel in the unprocessed image. These tests measured the absolute intensity of the pixel and the intensity relative to its immediate neighbours. The pixels that passed all three tests were then subjected to a region growing algorithm in order to establish the sets of connected pixels that represented individual microcalcifications. A second set of tests based on size and edge strength values was then applied to the identified microcalcifications and those that passed were subjected to a cluster filter to identify groups of three or more microcalcifications in a 1cm x 1cm area.

Fam applied this system to a set of 40 mammograms, digitised with a 0.2mm per pixel sampling rate, all of which contained clustered microcalcifications. The system found all of the clusters that had been identified by the radiologists as well as four clusters that had not been observed at the initial screening. Only two false–positive clusters were found in the 40 cases.

This system appears to have performed well, at least with the limited test set used, though Fam reports that the algorithm did require some manual adjustment to compensate for contrast and intensity variations. This need for manual control would be difficult to overcome without some sort of normalisation process to counteract the variations in image properties that could take the feature values of microcalcifications out of the characteristic ranges.

Chan and her colleagues (1987b, 1988, 1990) have developed an automatic detection system that employs a certain amount of image processing prior to feature analysis. Chan's algorithm involved processing a digitised image to yield a signal–enhanced image and a signal–suppressed image. Subtracting the two gave a difference image that was thresholded and processed by feature analysis.

Signal enhancement was achieved by means of a matched filter, a simple spatial filter that increased the peak intensity values of microcalcification pixels relative to background pixels.

Several different forms of filter were investigated for the signal suppression, with the most effective being the box–rim filter. The box–rim filter is an averaging filter with the weights of the central region set to zero. This has the effect of replacing a signal with the average value of the background. This signal–suppressed image was then subtracted from the signal–enhanced image and the result was segmented by local thresholding.

The structures present in the segmented image were then tested for suitable area and contrast values before a clustering principle was applied to locate groups of 3 or more microcalcifications in a 1.5cm x 1.5cm area.

The system was tested on a set of 20 clinical mammograms, digitised with a 0.1mm per pixel sampling rate. A true positive detection rate of 82% was attained with a false–positive rate of one cluster per image. By varying the signal–to–noise (SNR) ratio used to determine the local thresholds, the true positive rate could be increased to 100%, but at this value of SNR the system also detected 36 false–positive clusters per image.

While this system does not appear to exhibit as high a level of performance as that of Fam's algorithm, Chan's system did have the distinct advantage of being fully automated. It is possible that, had no manual adjustment been made in Fam's experiment, the performance of the system would have been significantly reduced.

Of course, it should be stressed that care must be taken when comparing the results of these and other systems. The limited sample sizes of the test data, and the observed variability in mammographic images could mean that such factors as the degree of subtlety of the microcalcifications could vary greatly between the different studies.

Spiesberger and Groh (1977) developed a calcification detection system that used feature tests based on contrast, brightness and compactness. The interesting thing about this system was that in order to reduce false–positives, two mammographic views of the same breast were processed and then correlated, eliminating the suspicious structures that only appeared in one view. Unfortunately the detection accuracy of the system was not reported, so it is difficult to assess how effective the procedure might have been. However, it does seem feasible that the comparison of different views could be applied to other systems in order to improve detection performance.

4.2.3 Feature Analysis

In order to use pre–selected feature values in systems such as those described above, it is necessary to establish the characteristic feature values of microcalcifications.

Olson and her team (1988) analysed 48 images containing 52 microcalcification clusters, the locations of which had been previously identified by a radiologist. Characteristic feature values were extracted for individual microcalcifications as well as for clusters; other variables such as the patient's age and type of mammography equipment used were also studied.

Olson used the feature values to compare benign and malignant clusters. She obtained a number of interesting results that seemed to suggest significant differences in the densities of the tissue surrounding benign and malignant

calcifications, rather than in the calcifications themselves. Of course, these results were based on a fairly small sample size and apply at a population, rather than at an individual, level. Nevertheless, studies of this type can serve to assist the targeting of tests used in detection systems at those features that are of greatest diagnostic significance.

Lanyi (1985) has also studied the characteristics of benign and malignant microcalcification clusters, concentrating mainly on the shape of the microcalcifications and the configuration of clusters. Using the information gained from this study, Lanyi developed a differential diagnostic system based solely on the cluster characteristics. The application of this system to 297 cases yielded a sensitivity of 97.6% and a specificity of 73.3%

Once again, this demonstrates that an understanding of the characteristic appearance of clustered microcalcifications can be a significant aid to their identification and classification.

4.2.4 Neural Networks

The success of pattern recognition systems in the detection of microcalcification clusters (see section 4.2.2) suggests that the task may be suitable for the application of neural networks, which have been used successfully for a number of pattern recognition and classification tasks (Rich & Knight 1991).

Bourrely and Muller (1990) have evaluated several variations of neural networks applied to the task of detecting clustered microcalcifications.

They used a neural network that was trained to classify input patterns as either microcalcification or background. The input patterns consisted of 20x20 pixel windows taken from digital mammograms. The network was trained with 200 patterns and then tested with 134 different patterns. The best results were obtained when the input pixels were logarithmically pre-processed. In this case

the network correctly classified around 90% of the microcalcifications and 70% of the background patterns, with the remainder consisting mainly of unclassified patterns rather than false-positives.

Bourrely and Muller concluded that the combination of logarithmic pre-processing and shared weights for the connections between layers of the network would result in an improvement in the system performance, though this combination was not tested.

The results of this study are encouraging, but it should be noted that the input pattern consisted of only a very small portion of a mammogram. There are approximately 2600 20x20 pixel windows in a 1024x1024 pixel digital mammogram. The processing of an entire image would therefore require either a substantially larger network or a considerable number of individual runs of the system.

4.3 The Detection of Lesions

4.3.1 Feature Testing

In a similar manner to their use in the detection of microcalcifications feature tests may be used to discriminate between lesions and normal breast structures.

Kegelmeyer (1992) used feature tests as the basis of an algorithm for the detection of stellate lesions. The algorithm operates on each pixel in the image and extracts five feature values including four Laws texture energy measures and an index of the distribution of local edge orientations. These tests form the nodes of a binary decision tree, so that each node is effectively a threshold on one of the extracted feature values, with the thresholds being established from training data. A pixel must pass all five thresholds to be accepted as part of a suspicious region.

The system was tested with 50 normal and 12 abnormal images. Half of these images were used to train the system while the others were used to test it. A true–positive rate of 100% was reported with 0.27 false–positives per image. At a lower operating point with a 92% true–positive rate there were no false–positives. Although these results are extremely encouraging, it should be noted that the number of abnormal images used for testing was very low.

Although Kegelmeyer used feature testing as the basis of his system, tests of this sort are more often used in addition to other methods as a means of improving the specificity of the system, in which case they usually occur in the latter stages of processing.

Giger and her colleagues (Giger et al 1990a & b, Nishikawa et al 1993, Yin et al 1991,1993) use bilateral subtraction of the left and right breasts to identify asymmetries that may correspond to lesions then use feature testing to refine the set of candidate locations.

In Giger's system each image is first thresholded ten times at different levels ranging from 5% to 50% of the area under the grey level histogram. Bilateral subtraction is then performed at each threshold level. Run length analysis is performed to find pixels that persist with a non–zero value in a series of more than five of the subtracted images. The final stage then involves the application of feature tests for such characteristics as size, circularity and contrast. This is a typical way in which feature testing is used after the bulk of the image processing has been performed.

Various results for the system have been reported. A test with 154 mammogram pairs (90 masses) showed a true–positive rate of 85% with three false–positives per image (Nishikawa et al 1993). A different study by the same group reported a true–positive rate of 95% with 3 false–positives per pair when tested with a set of 23 mammogram pairs, 18 of which contained lesions.

An interesting distinction between different applications of feature analysis arises from the types of features used, or more precisely from the method of selecting those features. A good feature test may be defined as one that has maximum discriminatory power for distinguishing between targets and non–targets. Unfortunately, due to the extensive variation in the appearance of lesions, there appears to be no single feature that can distinguish any lesion from any non–lesion which is why most systems tend to use a set of feature tests rather than a single test. There seem to be two methods for selecting which feature tests to use; some researchers choose a large number of feature tests in a fairly arbitrary manner and perform experiments to determine which of these tests are the most discriminating, while others make use of expert knowledge.

As an example of using expert knowledge, Giger (1990a & b) uses a simple test for classifying lesions as benign or malignant. This test involves smoothing the lesion contour to remove any spiculations then comparing the lesion before and after smoothing to determine the degree of spiculation, with a high measure indicating malignancy. This test was motivated by the fact that the degree of spiculation is an important diagnostic cue that the radiologist uses to determine malignancy.

One limitation of relying on expert knowledge is that the system is tied by the limitations of the human visual system. For example, Miller & Astley (1993) have found that Law's texture energy is quite good at distinguishing between glandular and fatty tissue – although, it is unlikely that the unaided human brain could explicitly calculate a measure of this type.

An alternative to using expert knowledge is to experiment with a number of statistically derived descriptors for measuring such characteristics as texture, shape or intensity gradients. Hoyer and Spiesberger (1979), for example, used ten statistical measures including texture measures derived from run–length analysis and second order statistics taken from the co–occurrence matrix. Semmlow (1980)

tested 19 measures for roughness, shape and spiculation, finally selecting nine of the most discriminating for use in the system.

Most commonly, feature tests are chosen on the basis of a grounding in expert knowledge and discriminatory power that has been proven experimentally. These common features include size, intensity gradients, contrast and simple shape descriptors such as circularity and eccentricity.

Feature tests seem to be most useful when used to refine a set of candidate locations that have been extracted from the image by some other processing technique. This not only improves the specificity of the system, but also reduces the computational cost by limiting the amount of the image that needs to be processed by these tests. Tests for fairly simple features such as size, circularity, contrast etc. have the advantage of being based on expert knowledge, which allows a certain amount of confidence that their discriminatory power will remain relatively constant over large numbers of images. In addition, they are fairly easy to implement. More sophisticated features involving the analysis of texture or second order statistics may appear to be effective for one group of images, but extensive testing on large data sets would be required before they could be accepted as reliable tests.

4.3.2 Asymmetry Detection

One property of mammograms that has been exploited by a number of researchers for the detection of lesions is that the images of the left and right breasts tend to be fairly symmetric. This means that a lesion in one breast should be detectable as an asymmetry if the breasts are compared.

Unfortunately, the left and right breast images are not perfectly symmetric. Apart from natural structural asymmetries in the breasts themselves, the mammographic imaging procedure involves compression of the breasts which can lead to

differential distortion of the images. In addition, variations in the imaging process can mean different distributions of intensities in the two images. All of these phenomena lead to enough minor asymmetries between the left and right mammograms to defeat any attempts at a straightforward comparison.

Nevertheless, bilateral comparison is an important mechanism used by radiologists when searching for abnormalities, and the use of asymmetry cues could help to improve the effectiveness of automatic detection systems. Yin et al (1993) have compared their method for detecting asymmetry by bilateral subtraction (see previous section) with a lesion detection algorithm that operates on only a single mammographic view. They found that their bilateral technique exhibited significantly higher detection performance than the single view method. However, it should be noted that the single view technique used was really rather unsophisticated and that other systems working on single views have shown superior detection performance (eg. Kegelmeyer 1992, Lai et al 1988, 1989).

In general, most researchers investigating asymmetry detection have assumed a spatial correspondence between the two breast images and have attempted to align the breast images as much as possible before comparison. Hoyer and Spiesberger, for example, divide each of the breast regions into an equal number of rectangles, extract a number statistical measures from each rectangular patch and then compare the corresponding rectangles from the left and right breasts to find any asymmetries.

Miller & Astley (1993) however, take the view that a comparison of anatomically corresponding regions is more appropriate than looking at spatial correspondence. Their studies have demonstrated that radiologists can still achieve a relatively high level of lesion detection performance (true–positive rate of around 70%) even when all diagnostic cues have been removed except the shape of the non–fat regions of the image.

Their system operated by segmenting out the glandular (non–fat) regions in each breast using texture measures. A number of measures were then calculated for each region and compared to determine whether the regions were asymmetrical. With the exception of transportation all of the measures used were global rather than local, so there was no need to align the regions.

Three types of asymmetry measure were used; shape (compactness, circularity, eccentricity and fourier features), brightness distributions (moments and transportation) and topology (area and binary moment). Of these features, experimentation revealed six with the greatest discriminatory power.

The system was tested with 104 mammogram pairs, each of which contained a single asymmetrical abnormality such as a mass lesion, an architectural distortion or a focal density. A true–positive rate of 67% was reported with a false–positive classification rate of 17%. Overall 74% of the pairs were correctly classified.

In most of the asymmetry detection systems, one of the first steps is to align the left and right images. Methods of image alignment vary greatly in their complexity. For example, Semmlow (1980) uses a fairly simple method that involves alignment to match up the extreme points on the image contours (the nipples) and then alignment in an orthogonal direction by application of a least–square–error technique to the differences between the two borders. Giger simply superimposes the breast contours and uses the smallest common area for further processing, though more recently rotation and translation procedures have been added to this basic method.

Simple alignment procedures as used by Semmlow and Giger will inevitably lead to a certain amount of correspondence error. Yin (1991) has suggested that most masses are large enough to not be affected by slight misalignments of the breast images. This may be the case, but it is the smaller, subtler lesions which might be

affected by minor misalignments and these should be the targets of an automatic detection system.

Lau and Bischof (1991) use a much more sophisticated set of procedures including rotation, translation, skewing and scaling to account for differences in orientation, position, shape and size between the two images. The result of these procedures are two breast images that have an exact spatial correspondence on a pixel–for–pixel basis. This seems like a promising method, since it should avoid artifacts caused by contour misalignment. Giger, for example, had to include a 'border test' in her final analysis to remove any suspicious locations that were associated with the border contour.

As Miller and Astley point out, the problem with a method like Lau and Bischof's is that it involves a certain amount of distortion of the information in the mammogram, so that the image no longer gives a true representation of the breast structure. Additional problems could arise for a prompting system, since the lesion located on the distorted image would quite possibly have a different location on the original – requiring that the detected locations be mapped back through the alignment procedures to the original image.

The actual comparison of images has been done in a number of ways, which fall into two groups. The first group of methods involve bilateral subtraction, either of the original grey level images (eg Giger 1990a–c, Yin 1991, 1993) or of versions that have received a certain amount of processing (eg Semmlow 1980, Lau & Bischof 1991). These methods have the advantage of being fairly easy to implement and of quickly eliminating a large proportion of the non–target information. However, these methods are extremely vulnerable to small non–target asymmetries that do not represent lesions and are generally used in conjunction with some sort of feature testing to improve the systems specificity.

The second group of methods involve extracting some feature value such as texture or brightness distribution from the two images and comparing the values

either globally (eg. Miller & Astley 1993) or on a more local level (eg Hoyer & Spiesberger 1979). These methods are generally less vulnerable to small non-target asymmetries but do involve all of the problems associated with feature analysis.

4.3.3 Multi-resolution Analysis

Another group of detection techniques take advantage of the fact that lesions tend to persist (remain visible in the image) at a variety of scales, ie; when the resolution of the image is reduced.

The most significant of these studies is the work of Brzakovic (1990) which was based on the technique of 'fuzzy pyramid linking'. Firstly, a gaussian pyramid was constructed such that each layer was half the dimensions of the layer below it and every pixel was the weighted average of a 4x4 pixel window in the level below. The weights consisted of a gaussian mask. Each pixel was then linked to its four candidate 'father' pixels on the level above it with a link that had a strength which derived from a fuzzy membership function applied to the difference in grey level between the father and son pixels. Once each pixel had been linked, the values of every pixel on all the layers above the base were updated to be the weighted average of that pixel's sons. This process of linking and updating was repeated iteratively until a steady state was reached, then the image was segmented by propagating the values of high level nodes back down through the pyramid.

Brzakovic's method has a lot of intuitive appeal. Firstly, the technique does not involve any *a priori* knowledge of the image or lesion characteristics, (although Brzakovic does use some simple feature testing at a late stage to improve the false-positive rate), which means that in theory the method should be able to detect any lesion regardless of size or shape. Secondly, the method, in fact the whole pyramid, can be used to detect microcalcifications as well as lesions with

a fair degree of success (Brzakovic, 1993), which suggests that it might be the basis of a general abnormality detection system.

Brzakovic's system correctly classified 85% of 15 films containing 10 mass lesions and demonstrated a false–positive classification rate of zero. Testing with 67 images including 17 containing microcalcifications showed a true positive detection rate of about 88% again with no false positives.

Semmlow (1980) also took advantage of the scale persistence of lesions and constructed a series of reduced size primary resolution cell (PRC) images, where each pixel (PRC) represents a property of a 10x10 pixel window in the original image. One of these PRC images was based on average intensity, and was therefore really just a reduced resolution image. Other PRC formats including intensity gradients, roughness and shape descriptors. In this case, however, the PRC images were used to generate asymmetry cues and subjected to feature analysis, rather than being used as the basis of the detection algorithm as in Brzakovic's system.

4.3.4 Other Methods

Two other methods of lesion detection that are worth noting but do not fall into any of the previous categories are template matching (Lai et al 1988, 1989) and neural networks (Nishikawa et al 1993).

The template matching method developed by Lai uses the observation that well–defined lesions are generally circular or near circular in shape. The first stage of this algorithm involves noise suppression by the application of a median filter to the image. In this case the median filter was specially modified with a threshold that served to preserve edge information. Lai compared what they called the "selective median filter" (SMF) with a number of other edge–preserving noise suppression algorithms and concluded that the SMF was the most effective.

The filtered image was then cross–correlated with a series of circular templates ranging in radius from 2 to 14 pixels. Each of the templates consisted of 1's inside the circle and –1's outside the circle with zeros at the boundary. These zeros constituted a free area that allowed for deviations of the lesion from a perfect circular shape. The top 2.5% of the cross–correlation measures were then thresholded out as suspicious regions. Finally two feature tests were applied to improve the system specificity.

The system was tested on a rather small data set (19 tumours in 17 images) and showed a 100% true–positive rate with 1.7 false–positives per image.

Nishikawa and colleagues (1993) described the application of a neural network to the classification of lesions. They identified 14 characteristic features of the signs of early breast cancer. Five of these related to lesions (eg degree of spiculation), six were related to associated microcalcifications and three were related to secondary features such as skin thickening. These features were applied to a set of lesions and and for each lesion, a value for each feature was assigned by a radiologist. These values were then fed into a neural network with a single output node in order to classify the lesions as benign or malignant.

The network was trained on 60 clinical cases using the leave–one–out technique. In addition to the training cases, 133 textbook examples of lesions were used to test the system. The classification performance of the network was reported to be slightly better than that of a radiologist used for comparison – though no figures are given for the results. At present the manual extraction of feature values makes this system far from automatic, and the performance of the network would be expected to decline rapidly if it had to perform the additional task of automatic feature extraction.

4.4 Summary – The Detection of Abnormalities

The preceding sections have described a number of systems that have been developed for the detection of abnormalities in digital mammograms, some of which have been more effective than others. The relatively high detection performance of some of these algorithms, especially in the detection of microcalcifications suggests that continuing development will lead to highly accurate systems, although, at present the development of a fully automated system for pre–screening mammograms is still a rather distant goal.

At present the most feasible application of these detection systems is as computer–based aids to assist the radiologist in the screening task. This observation has a number of important implications. Firstly, for the computer–based aid to be effective, the results of the detection system must be presented to the radiologist. This raises questions concerning human perceptual and attentional processes as well as human–computer interaction, and some of these questions have been addressed elsewhere in this report.

Another important implication that affects the development of CAD systems concerns the targeting of abnormalities. Ideally a CAD system would assist the radiologist by finding and drawing attention to abnormalities that the radiologist would otherwise have missed. The factors that lead to mammographic abnormalities being missed by radiologists have not really been studied in any depth, and certainly no CAD system has been developed on the basis of such data.

Research into the development of automated detection systems has focussed mainly on improving the detection accuracy of the system in its own right. While this is a worthwhile goal, it may not be the most appropriate direction for the development of computer–based aids to detection. It may be the case that an effective CAD system would exhibit a lower detection performance, in an

absolute sense, than an automated detection system, but be more useful to the radiologist because it locates abnormalities that might otherwise have been overlooked.

Chapter 5

Implementations of Prompting Systems

In order to assess the effects of prompting on the detection performance of radiologists a system was required for generating prompts. This section will describe in detail two such systems. The first is based on an algorithm developed by Astley and Taylor (1990) and is specifically targeted at microcalcifications. The second is based on the 'fuzzy pyramid' developed by Brzakovic (1990) for the detection of both microcalcifications and lesions. In addition to explanations of these systems, results are presented that compare the performance of the two algorithms for detecting microcalcifications and assess the performance of the latter system for detecting lesions.

5.1 Combining Cues

5.1.1 Overview

The first of the algorithms involves the combination of evidence from two cue generators, both of which used grey level morphology. Each of the cue generators was selected to respond to a particular property of microcalcifications: that they have relatively sharp edges, and that they appear as small bright blobs. Figure 5.1 shows a diagrammatic overview of the algorithm.

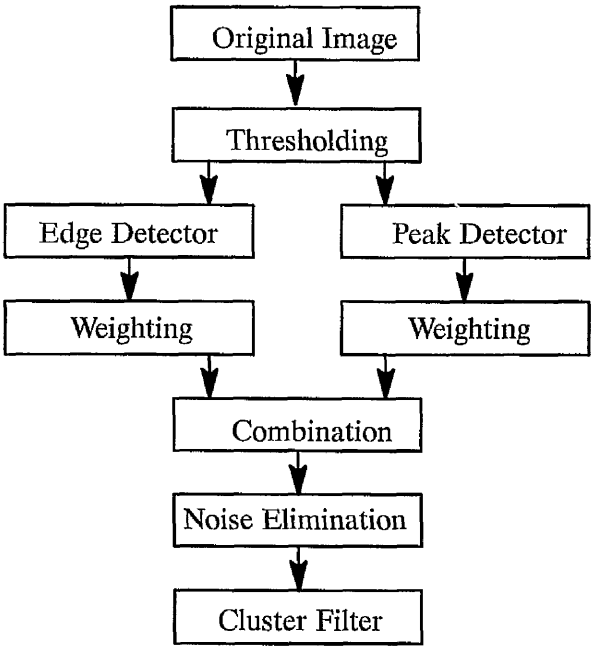

Figure 5.1: Diagrammatic overview of the cue combination method

5.1.2 Description of Algorithm

It was noted in testing the system that the non-breast regions of the original images were far from uniform and the cue detectors tended to pick up small discontinuities in these areas. In order to avoid this problem a simple global threshold was used to set the lowest 5% of the grey level histogram to zero, which effectively eliminated any variation in the darker non-breast regions with little effect on the breast tissue.

The thresholded image was then processed separately by each cue generator. The first cue generator, a morphological edge detector, consisted of an eroded image subtracted from the original (thresholded) image and should have picked up the sharp edges of the microcalcifications. The second cue generator was a morphological top hat transform which involved performing a large scale closing operation and subtracting the results from the initial image. In this case the closing

operation involved 5 passes of erosion followed by 5 passes of dilation, which should have found bright objects up to 10 pixels (1 mm) in diameter. The top hat transform was preceded by a standard single pass closing operation in order to remove very small noise objects (<0.1mm diameter).

Each of the cue generators should have picked up the microcalcifications as well as certain other structures with similar properties. However, the microcalcifications were rarely the most prominent structures in the cue generator images and it was difficult to differentiate between targets and non–targets. It was therefore necessary to weight the results of the cue generators to selectively enhance the microcalcifications.

The weighting procedure was based on a simple statistical model of microcalcifications, or more specifically a model of the typical responses of the cue generators to the microcalcifications. The model was obtained by a study of 594 microcalcifications that had been located and marked by a radiologist. Each microcalcification was marked in the (approximate) centre and in order to locate the edge a region growing algorithm was used. This algorithm involved repeatedly dilating the central marked pixel and comparing the results with the original image. A pixel could only be 'claimed' by the dilation if its grey level on the original image was within 15% of that of the pixel being dilated. The value of 15% allowed for a reasonable amount of variation in grey level within the microcalcification. It was found that 2–4 passes of dilation were generally required to 'fill in' the microcalcifications, and if the number of passes was greater than 5 then it generally meant that the dilation had 'leaked' beyond the boundary of the microcalcification. The 23 cases in which this 'leakage' occurred were excluded from the results. A morphological inner–edge detector was then used to locate the edge of each microcalcification.

Each of the cue generators was then applied to the locations containing the known microcalcifications in the original images. In the case of the inner–edge detector

the resulting values of each point on the edge were averaged to give a single value for each microcalcification, while the top–hat used the average value of all of the pixels within the microcalcification. This procedure yielded an average response value for each of the two cue generators applied to each of the 571 microcalcifications in the training set. By analysing the distributions of the responses of each of the cue generators it was possible to calculate the mean and standard deviation in each case. These statistics formed the basis of the weighting procedure.

Figure 5.2 illustrates the function used to weight the cue generator responses. The results of each cue generator were weighted separately, using the appropriate statistics taken from the distribution of that generator's typical responses to microcalcifications, though the weighting procedure was the same in each case.

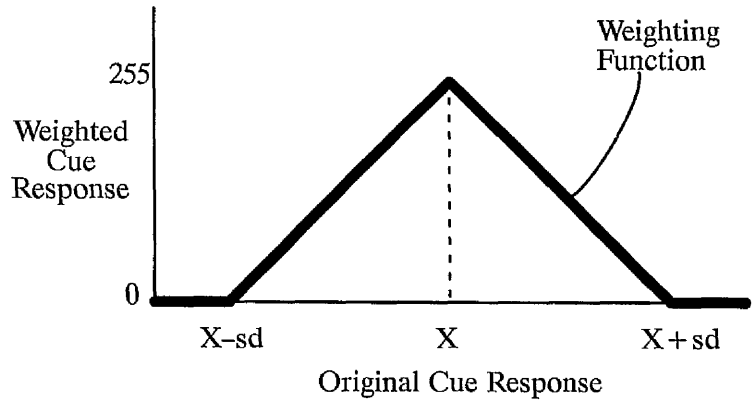

Figure 5.2: Function used to weight cue images (X = mean, sd = standard deviation)

In the weighted cue image those points in the original cue image with a grey level equal to the mean of the typical responses for that cue generator were given the maximum grey scale value (255). The weighted value assigned to other points was inversely proportional to the difference between the original cue image value and

the mean of the typical responses. When this difference exceeded a certain multiple of standard deviations of the distribution of typical responses the rescaled value was set to zero (background). The value of the multiple of standard deviations was varied in order to achieve different levels of response bias for the system.

The probability that a pixel was part of a microcalcification was now represented implicitly by its grey scale value in the weighted cue images and many of the non–target structures, such as the breast boundary, which had generated very strong edge cues were eliminated from the weighted images since they fell out of the range of the weighting function.

Although each cue generator responded to the target microcalcifications, each of them also responded to a number of other structures in the images that shared certain properties with microcalcifications. In order to improve the overall specificity of the system the evidence from the two cue generators was combined, with the intention of suppressing potential targets with evidence from only one generator and enhancing potential targets with evidence from both. The method used to combine the evidence was a straightforward multiplication of the two weighted cue images, which was particularly effective after the weighting procedure since the grey level of each pixel in the weighted image was an implicit representation of the likelihood that the pixel was part of a microcalcification.

At this stage the microcalcifications were clearly present in most images, however there was generally a certain amount of noise also present – usually very small single or double pixel blobs. Initial attempts to eliminate this noise by means of a closing operation had the effect of removing too much target information, so an erosion procedure was used with an additional constraint that prevented the removal of a pixel if it was connected to at least two other pixels. Multiple passes

of this procedure were applied until no more of the small noise blobs remained, with an average of 6 passes per image.

The final stage of the system involved searching the remaining objects in the system to see if any fell into clusters in the manner of microcalcifications. In order to achieve this, a mask of fixed size, 100x100 pixels ($1cm^2$), was applied to each blob in the final image. In each case the mask was applied in each of the four positions that included the target blob in one of the four corners of the mask. At each position the mask was checked to see whether sufficient microcalcifications lay within its boundary to constitute a cluster. A minimum of 3 blobs were required for a cluster to be accepted. If the mask located a cluster, then the locations of all the constituent microcalcifications were averaged to obtain an estimate of the cluster centre which was recorded and used to generate a prompt.

4.5.3 Typical appearance at different stages

Figure 5.3a–d illustrates the various stages of the morphological algorithm, beginning with a patch taken from a digital mammogram that shows a cluster (Figure 5.3a). Figures 5.3b and 5.3c then show the output of the weighted edge detector and the weighted top hat respectively, and finally figure 5.3d shows the final result after the two cue images have been combined and small non–target blobs have been filtered out.

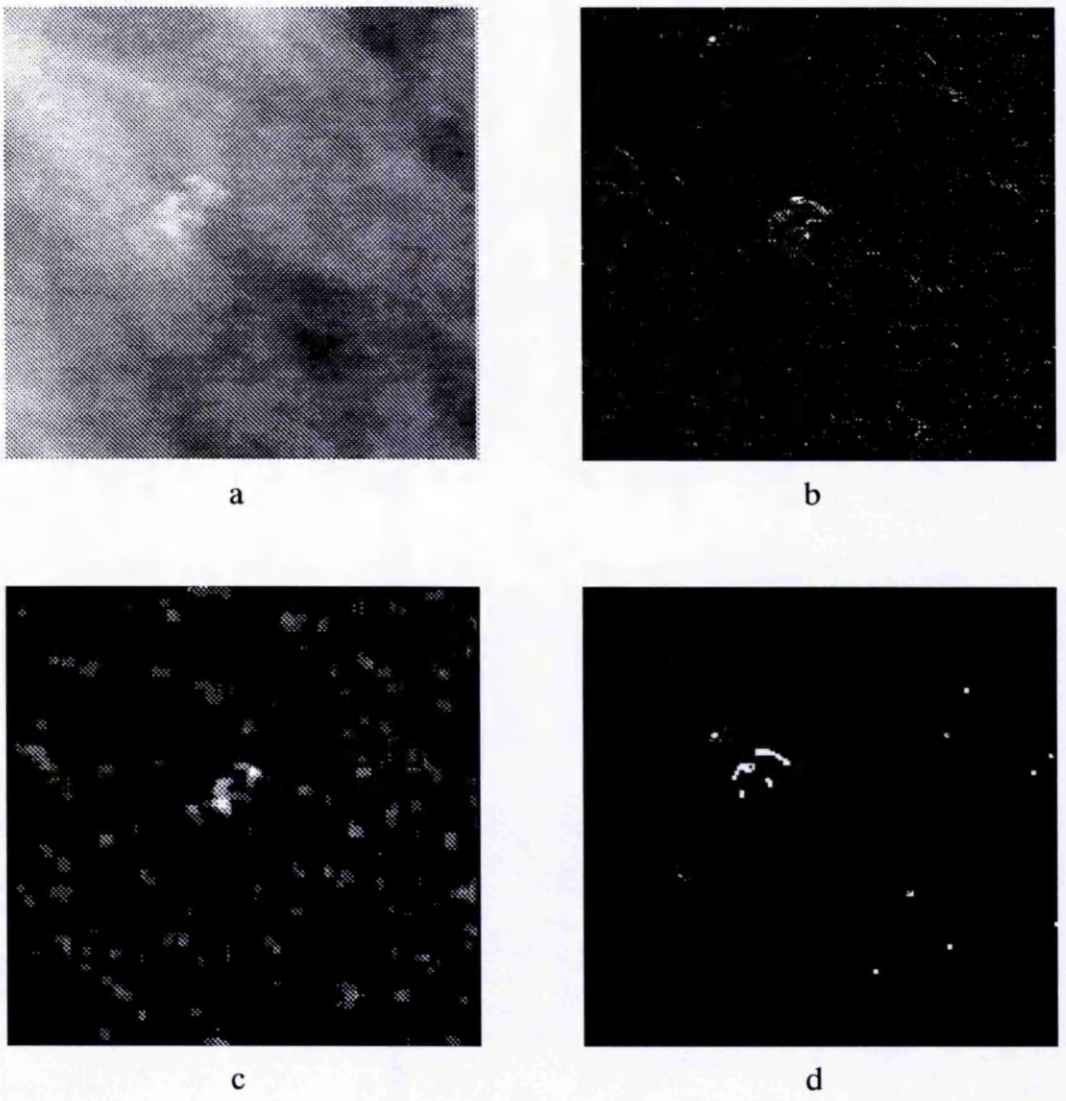

Figure 5.3: Stages in the detection of a cluster:
a. source image, b. weighted edge cue image,
c. weighted top hat image, d. combined image

Figure 5.3 illustrates the typical appearance of a single cluster at the various stages involved in its detection. A fuller discussion of the results of the algorithm applied to a large test set of clusters will be presented in section 5.3.

5.2 Fuzzy Pyramid Linking

5.2.1 Introduction

The second method used for the detection of microcalcifications was a slightly modified version of the fuzzy pyramid linking algorithm described by Brzakovic (1990). One advantage of this method is that very little *a priori* knowledge about microcalcifications is required until the final stages of the algorithm, which involves testing the features of any structures passed by the segmentation procedure. This suggests that the algorithm could also be used for the detection of abnormalities other than microcalcifications.

5.2.2 Constructing the Pyramid

Prior to construction of the gaussian pyramid, the original image was pre–processed by contrast enhancement which involved simply 'stretching' the grey level histogram to occupy the full range of available grey levels. The images used in testing our system were 512x512 pixel patches extracted from mammograms, and as a consequence often only employed a part of the grey level range. Since the fuzzy linking procedure used in the algorithm operated on the difference in intensity between pixels, the combination of contrast stretching followed by the gaussian smoothing that was a natural part of the pyramid construction allowed for a greater range of link strengths than would otherwise have been the case, facilitating the discrimination of targets from non–targets on the basis of the link strengths.

The next step in this method was the construction of a gaussian pyramid with the (enhanced) original image as the base, as illustrated in figure 5.4. Each level of the pyramid above the base (level 0) had half the dimensions of the level below it, with the value of each pixel in a level above the base being generated by the application of a 4x4 gaussian mask to the pixels on the level directly below it.

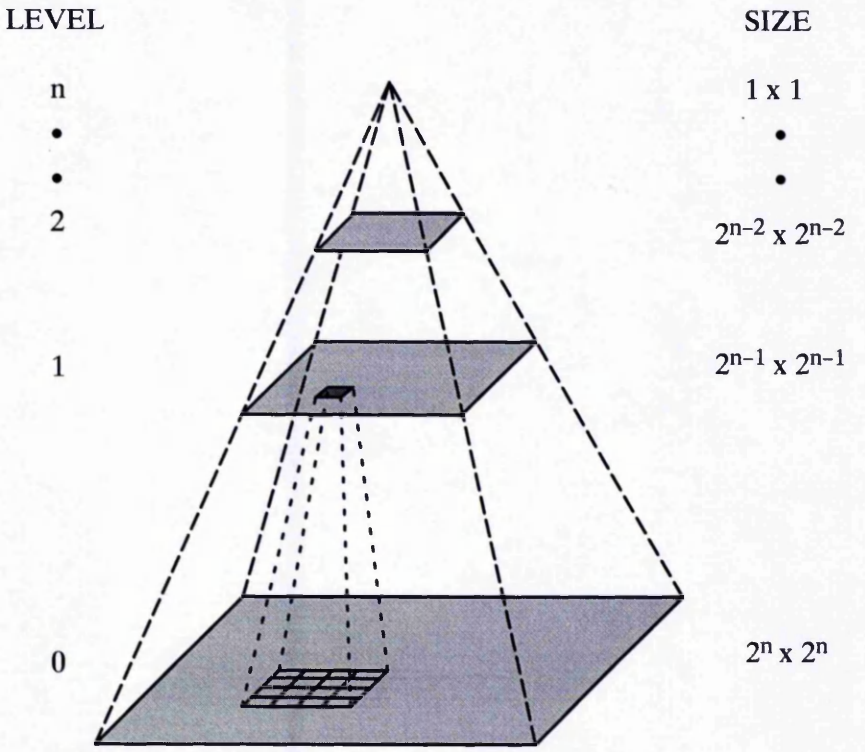

Figure 5.4: Structure of gaussian pyramid

5.2.3 Fuzzy Linking

Once the pyramid had been built the levels were linked together. Each node (pixel) on a level above the base had 16 possible son nodes on the level below that could be linked with it. Similarly, each node in a level below the top two had four possible father nodes on the level above that it could link with.

The simplest method of linking the levels together would have been to use 'hard linking' as described by Burt (1984). Using hard linking the four candidate father nodes are compared and the one closest in grey level to the son is linked, while the other three are not. Every son node is linked to a father in this way, and many fathers have more than one son. However, rather than this hard linking procedure, fuzzy linking as described by Brzakovic was employed in this algorithm.

The fuzzy linking method allows any given node to be linked to all of its potential father nodes and all of its potential son nodes, with each link having a link strength determined by a fuzzy membership function (Zadeh 1965) operating on the difference in grey level between two linked nodes. Each node is linked with four father nodes and the link out of these four that has the highest link strength was known as the maximum link for that node.

The equations that determine the strength of the link between two nodes (Φ) can be expressed as follows;

$$\Phi_{i,j,i',j'}(u;\ \alpha,\ \beta,\ \gamma) = 1 - S(u;\ \alpha,\ \beta,\ \gamma),$$

where:

$$S(u;\ \alpha,\ \beta,\ \gamma) = \begin{cases} 0 & \text{for } u <= \alpha \\ 2\left(\frac{u-\alpha}{\gamma-\alpha}\right)^2 & \text{for } \alpha <= u <= \beta \\ 1 - 2\left(\frac{u-\gamma}{\gamma-\alpha}\right)^2 & \text{for } \beta <= u <= \gamma \\ 1 & \text{for } u >= \gamma \end{cases}$$

$$u = |I_L(i, j) - I_{L-1}(i', j')|$$

$I_L(i, j)$ = intensity of node at point (i, j) in level L of the pyramid

α and γ are parameters that determine the shape of the function

$$\beta = \frac{\alpha + \gamma}{2}$$

The shape of the fuzzy membership function used to determine link strengths is defined by two parameters; α and γ. Brzakovic's implementation of this algorithm used fixed values for the parameters α and γ. However, it was noted during testing of the algorithm that using fixed values for these parameters did not account for the degree of grey–level variation observed in the original images. This led to a

reduction in the discriminating power of the algorithm as the full range of possible link strengths was not being used. To overcome this problem α was fixed at zero and γ was automatically selected to be one standard deviation of the grey level distribution of the (contrast enhanced) original image. This meant that the full range of possible link strengths was used.

Once all the link strengths at a given level had been established, the values of the nodes at that level were updated. This was achieved by replacing the value of any given node with the weighted average of the son nodes that were linked with it from below. The weight associated with each son node corresponded to the strength of the link between that node and the father node.

Once all of the nodes had been updated the process of linking and updating was repeated continuously, forming an iterative process. At each repetition, the number of nodes that had their maximum link with a different father than on the previous iteration was counted. Once this number reached zero, the pyramid had converged to a steady state, and there were no more passes of linking and updating.

The original image was now segmented by taking the values at the 2x2 level of the pyramid, just below the apex, and propagating these values down through the pyramid along every link that exceeded a certain threshold on the link strength. The level of this threshold was varied to provide a number of different levels of response bias and allowed an ROC curve to be generated.

The final stage involved the application of feature tests based on the properties of the abnormality being searched for. In the case of microcalcifications this involved searching the image for clusters of 3 or more blobs with maximum nearest neighbour distances of 5mm (50 pixels). For lesions, the image was subjected to morphological opening then searched for objects with a minimum area of 25 pixels.

5.3 Results

5.3.1 Comparison of Methods – Microcalcification Detection

In order to assess the effectiveness of the two detection algorithms described in the preceding sections, each was tested with the same group of 60 images. The images were 512x512 pixel patches that had been taken from digital mammograms with a spatial resolution of 10 pixels mm^{-1}. Of these 60 patches 36 contained at least one cluster and three of these contained two distinct clusters – giving a total of 39 clusters in the data set. The remaining 24 images contained no abnormalities. The images were read by a radiologist who also had access to the original films, and the locations of any clusters were determined. In the case of the first algorithm, films were also required for training the system. In these cases, the 'leave–one–out' method was used to train and test the algorithm.

In order to generate points for an ROC curve, the systems were required to operate at a number of different levels of response bias. In the first case, this was achieved by varying the width of the weighting function in terms of a multiple of the standard deviation (sd) ranging between 0.6sd and 1.1sd in steps of 0.1sd. In the second case different operating levels were achieved by varying the threshold on the link strengths between 0.1 and 0.9 in steps of 0.1.

For each image, the numbers of true–positives and false–positives at each level of response bias were determined and the true–positive rates and numbers of false–positives generated per image were calculated for each system. These data are illustrated by the FROC curves shown in figure 5.5.

In order to derive measures of sensitivity that could be subjected to statistical analysis, the FROC curves illustrated in figure 5.5 were converted into ROC curves as described in chapter 3. The resulting curves are shown in figure 5.6. The solid–line curves in figure 5.6 represent the best–fit ROC curves for the

experimental data, as calculated using the ROCFIT package. Analysis of these results revealed that the fuzzy pyramid system exhibited significantly higher classification accuracy than the morphology–based system, (t_{obs} = 4.17 , $p < 0.005$).

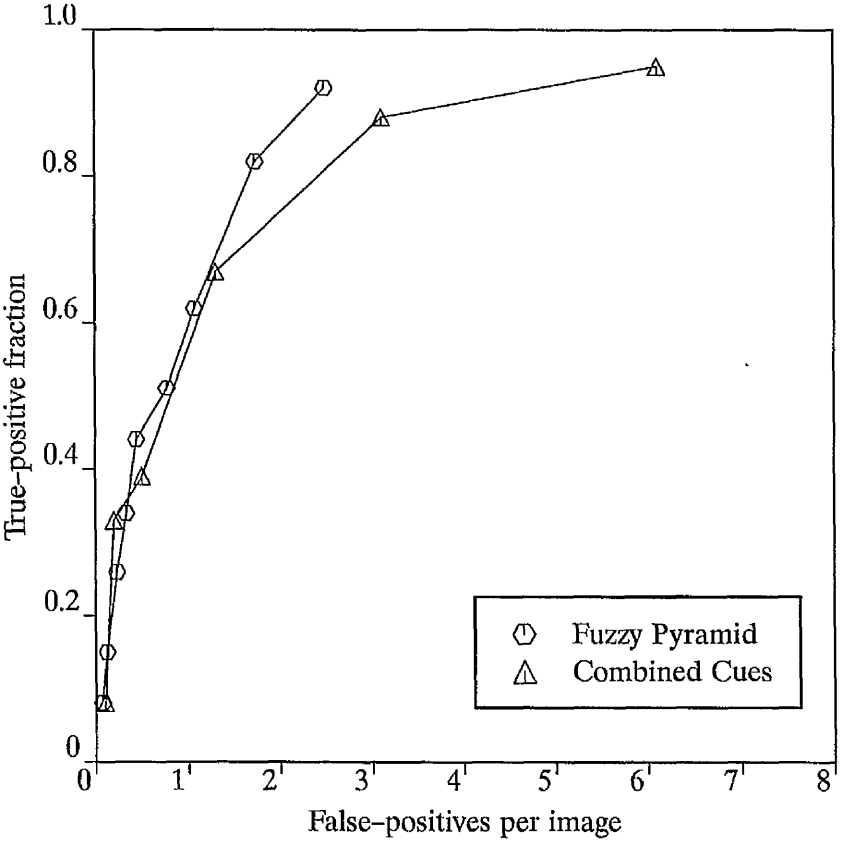

Figure 5.5: FROC curves showing detection performance

Clearly in this test the performance of the fuzzy pyramid algorithm exceeded that of the morphological system, with this increased performance manifesting itself as a lower number of false–positives generated at any given true–positive rate. However, in their present state of development neither of these algorithms seems to operate at a level of accuracy that would be suitable for a system to be used in a clinical environment. True–positive detection performance was encouragingly

high, reaching 92% in the fuzzy pyramid system and 95% in the morphological system, but the numbers of false positives generated at these operating points were approximately 2.5 per image and 6.1 per image respectively. Research on the effects of false–positive errors on prompting effectiveness has suggested that the benefit of the prompts as aids to the radiologist is diminished and may be lost altogether as the false–positive rate of the prompt generation system increases (see chapter 6). Even the 2.5 false–positives per image rate of the fuzzy–pyramid system may be too high for the prompts to be useful to the radiologist.

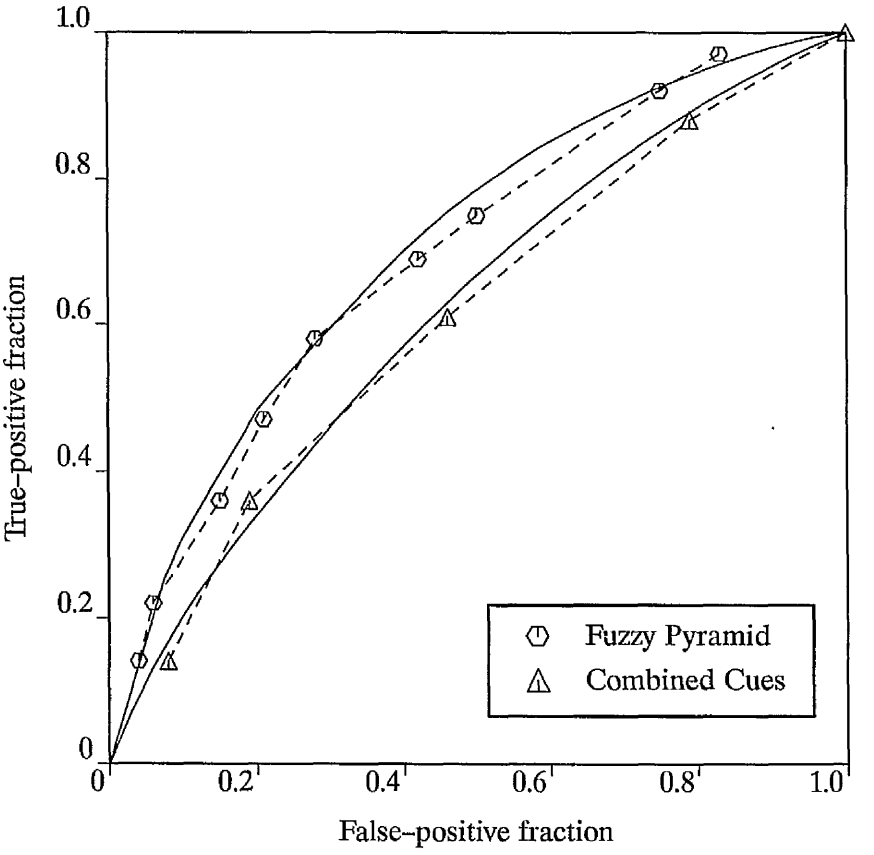

Figure 5.6: ROC curves illustrating classification performance

At present the final stages of processing in the fuzzy pyramid algorithm consist of simple tests to determine the size of potential microcalcifications in the final

image and whether or not they represent a cluster. By introducing some more sophisticated feature testing during these latter stages, it may be possible to improve the specificity of the system. For example it may be useful to determine the locations of any detected potential microcalcifications in the segmented image and examine these locations in the original image with a view to rejecting any that represent clearly normal tissue.

5.3.2 Lesion Detection Results

In addition to being applied to the detection of microcalcifications, the fuzzy pyramid system described in section 5.2 was also used to detect well-defined lesions.

The lesion detection version of the system operated in exactly the same way as described for the detection of microcalcifications, up to the thresholding procedure in the latter stages of processing. For the detection of microcalcifications, the threshold on the link strength served as an *upper* limit, so that only weak links were propagated down through the pyramid. This was the most effective method because the microcalcifications are small inconsistencies in the background pattern that lead to weaker link strengths. However, lesions take up larger areas of the image and have relatively smooth internal texture, which meant that the links associated with lesions were relatively strong, except at the edges of the lesion, where weak links were observed. For this reason, the threshold served as a *lower* limit in the detection of lesions.

The cluster detection procedure was clearly not appropriate for the detection of lesions and was replaced by an area threshold. The segmented image was first subjected to morphological opening (dilation followed by erosion) in order to 'fill in' the gaps that tended to appear within the lesions. Then any objects in the image with an area less than 25 pixels were rejected. The figure of 25 pixels was derived

from an empirical analysis of the typical size of lesions after segmentation by the pyramid algorithm.

The lesion detection system was applied to 15 images containing well-defined lesions, all of which had been digitised at a 0.1mm per pixel sampling rate. In order to generate an ROC curve, the link strength threshold was varied between 0.45 and 0.95 in steps of 0.05. The FROC curve in figure 5.7 shows the detection performance of the fuzzy pyramid algorithm in locating the lesions. For the purposes of comparison, the system's microcalcification detection performance ROC is also shown.

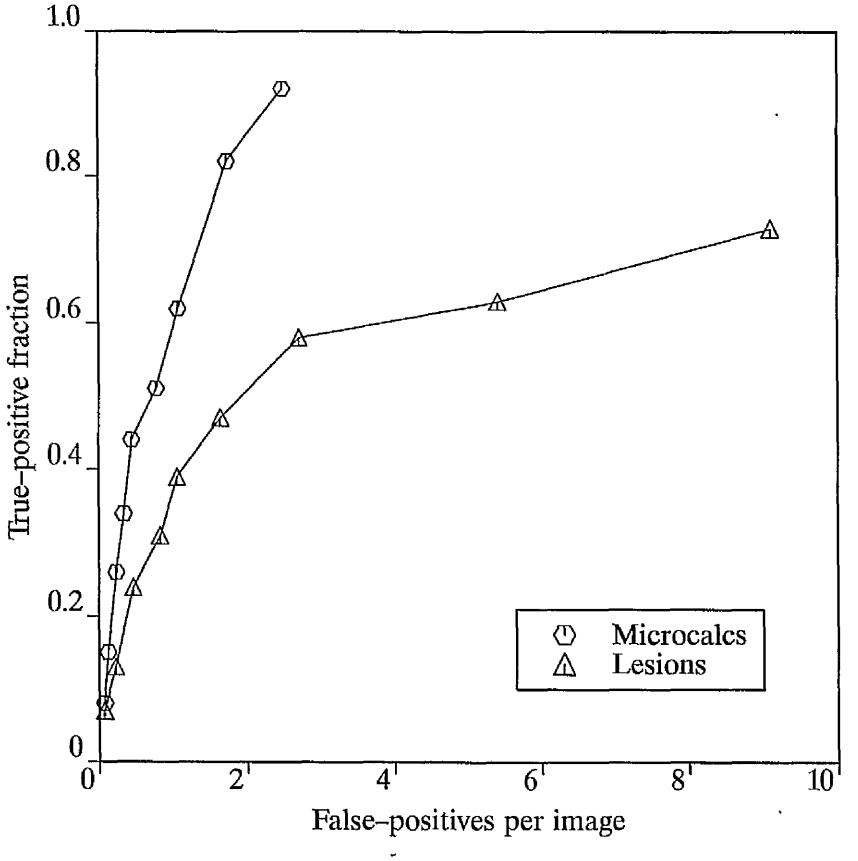

Figure 5.7: FROC curves showing detection performance of fuzzy pyramid system with lesions and microcalcifications

Clearly, the system is by no means as effective at detecting lesions as it is at detecting microcalcifications. This appears to be due to a tendency to classify any region of bright glandular tissue as an abnormality. Consequently the false–positive rate was rather high. In its current state the system does not take advantage of any of the properties of well–defined lesions that may improve the specificity of the system. Such properties may include edge strength, circularity, and contrast relative to the surrounding tissue. It is possible that by the inclusion of tests based on these properties may improve the performance of the system.

5.4 Summary & Conclusions

This chapter has described two systems that have been implemented for the task of detecting abnormalities in digital mammograms. It is not suggested that either of these systems represents the most effective or efficient method of detecting abnormalities of the types described. Rather, the intention was to generate realistic prompts for use in psychophysical studies with radiologists and will be referred to in subsequent chapters.

The advantages of this are clear. Simulated prompts, selected by human intervention, have a great potential for bias – especially in selecting the locations for false–positive prompts. Automatically generated false–positive prompts are not placed randomly, but rather will be sited at locations that have some characteristics in common with the abnormalities being detected. It may well be the case that such locations would be the most likely to cause a radiologist to make an error, making the interaction between the prompt and the human observer particularly interesting.

Of course, it should be noted that different detection algorithms will rely on different properties of abnormalities. The false–positives generated by one system will therefore differ from those generated by another.

Chapter 6

The Effects of False–Positive Prompts

This chapter describes the first of three experimental studies designed to investigate the effects of prompting on the detection performance of radiologists. This experiment examines the effectiveness of prompting at different false–positive rates.

6.1 Objectives

The aim of this experiment is to investigate the effects of varying the accuracy of the prompt generation system on the detection performance of the radiologist when searching for one particular class of abnormalities, clustered microcalcifications.

It is very unlikely that any automatic prompt generation system will ever be completely specific, so that a certain number of false–positive, or invalid, prompts will be produced. It is important to know how such invalid cues may affect the performance of the radiologist.

Although the primary measure of detection performance will be the signal detection measure of sensitivity, A_z, a number of other variables will be studied

including the reading time, the active use of prompts, subjective ratings of usefulness and the relationship between invalid (false positive) prompts and false positive judgements from the radiologist.

6.2 Experimental Method

6.2.1 Data

The experimental data consisted of 24 pairs of mammograms, each from a different patient. Each pair comprised mediolateral views of the left and right breasts and in each case one of the pair contained a cluster while the other was normal. These mammograms were digitised with a spatial resolution of 10 pixels mm^{-1} and an 8–bit grey resolution. Previous studies have suggested that glandular pattern type can affect the detectability of microcalcifications (Hutt 1992), but in this experiment the pairing of each abnormal image with a normal from the same patient and consequently with the same glandular pattern type should have stopped the pattern type from affecting the results.

A 1024x1024 image patch was extracted from each of the digitised mammograms for use in the experiment. The patches from abnormal images were selected to contain the clusters, while the normal image patches were taken from a breast area roughly corresponding to that used for their paired abnormals. In both cases selections were made to minimise the amount of non–breast background present in the image patches. In order to increase the size of the data set each image patch was copied and reflected about the y axis, effectively doubling the number of patches available.

The 96 image patches were randomly assigned to four experimental conditions representing different levels of prompting accuracy. Of the 24 patches in each condition half were normal and half contained clusters. With the exception of

those patches assigned to the control group, the image patches were all processed using the combined cues algorithm described in chapter 5. Each experimental condition represented a different combination of true–positive and false–positive prompting rates. These rates were obtained by running the detection algorithm at different operating points in a manner similar to that used to generate the ROC curves in section 5.3. Figure 6.1 summarises the accuracy data for each of the experimental conditions.

The 'level 2 accuracy' condition represents the true performance of the detection system with the width of the weighting function set at 1.1 standard deviations, which reflected an appropriate operating point as determined by ROC analysis. By running the algorithm with stricter criteria (narrower weighting functions), a true–positive rate of 60% and a false–positive rate of 0.5 image^{-1} were achieved. These were then combined with the 'level 2 accuracy' results as described in figure 6.1.

Condition	True Positive Rate	False Positive Rate
Level 1 accuracy	89%	0.5 image^{-1}
Level 2 accuracy	89%	2.4 image^{-1}
Level 3 accuracy	60%	2.4 image^{-1}
Unprompted (control)	—	—

Figure 6.1: Summary of experimental conditions

By combining true–positive and false–positive prompting rates from different operating points in this way, the three experimental conditions simulate three different systems, all with different levels of sensitivity.

The prompting rates used in the 'level 1' and 'level 3' conditions are artificial and do not represent the true performance of the detection system. It could be argued that for this reason it would have been easier to simply simulate the prompts.

However, simulating the prompts would have introduced a certain amount of bias as it would be necessary to somehow decide which abnormalities are prompted and which are not. A method for deciding where to place false–positives would also be required. As it stands, all of the prompts used in this study were genuine, in that they were all generated by a detection algorithm.

6.2.2 Subjects

The experiment was carried out with seven radiologists, three of whom were experienced in the reading of mammographic films and four of whom were involved in general radiological practice but were familiar with mammograms. None of the subjects had any significant experience in reading *digital* mammograms in a clinical setting.

Ideally, all of the subjects would have been experienced mammographers. However, the non–portability of the experiment restricted the number of mammographic radiologists available to participate.

Since the experimental task was rather different to conventional mammographic film reading (partial, digital mammograms presented on a workstation), any advantage associated with experience in reading mammograms should have been diminished.

6.2.3 Procedure

The images were presented to the subjects on a SUN Sparc workstation with a pixel size of 300 μm and a screen size of 1320 x 1035 pixels. Since the resolution of the images gave a 100 μm pixel size, the image patches were magnified by a factor of 3 in the experimental display. All of the subjects responses were made using a 3 button mouse and pad.

In order to maintain some level of consistency in lighting conditions, the experiment was carried out in a darkened room, though not in a total blackout.

The subjects were free to use the brightness control on the VDU in order to adjust the light level of the display and all of the subjects were observed to use this control frequently throughout the experiment. The layout of the screen display is illustrated in figure 6.2

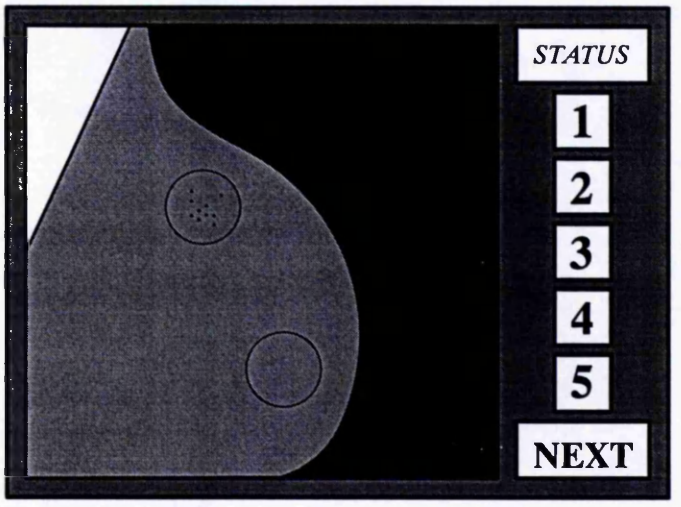

Figure 6.2: Diagram showing screen layout during experiment. A simulated cluster and two prompts are also shown, one TP and one FP.

The main portion of the display consists of the *image window* in which the image patches were displayed individually in sequence. At the upper right of the display is the *status window* which displayed brief instructions to the subject concerning their response options at any given point in the experiment. In a vertical column below the status window are 5 *response buttons* numbered 1 to 5. These were used by the subject to make confidence judgments. Finally, to the lower right of the display is the *next image button* which allowed the subject to move onto the next image in the sequence. All responses were made by clicking with the mouse on the appropriate button.

Each of the four experimental conditions was presented to the subjects as a separate block, with the order of block presentation randomised differently for

each subject. In addition the assignment of images to the various experimental conditions was different for each subject. In each case the assignment was random with the constraint that an image patch and its reflected version could not both appear in the same block. The order in which the images appeared within the block was also randomised. These precautions should have served to minimise or hopefully eliminate order effects arising from practice or fatigue.

Each of the images appeared twice in the study, once in its original state and once reflected, and therefore there is a possibility of learning influencing the results. However, the process of reflecting the image should have made it more difficult to recognise and when this is combined with the complexity of the image and the practice of separating the two versions, learning should not have been a major problem. In fact, when questioned after the experiment, none of the subjects reported having realised that the images were repeated.

Prior to the experiment, each subject was given verbal instructions concerning the nature and requirements of the task and the number of images in each block. In addition the subjects were informed of the proportion of abnormal images (50%) since this is much higher than the proportion that might be seen in a clinical setting such as a screening centre and the disparity between the expected proportion of abnormals and the actual proportion may have adversely affected the detection performance of the radiologists had they not been informed. Similarly, the subjects were informed of the accuracy of the prompting system before each block was presented. Kundel and Nodine (1978) have suggested that verbal instructions concerning the patient's clinical history and other salient factors can serve to modify the search behaviour of the radiologist searching for small lung abnormalities. It is possible that informing the subjects of the prompt accuracy may have affected their search strategy slightly differently for each experimental condition in this experiment, but it is almost certainly the case that the subjects would have determined the approximate accuracy of the prompts after a few

images, at which point they would have adjusted their search strategy anyway. By informing the subjects at the outset their search strategies should have remained fairly consistent throughout each block.

In addition to verbal instructions the subjects were also given a practice run in order to familiarise them with the display and the appropriate responses. The practice run consisted of 6 images representing a cross–section of the experimental conditions and subjects were invited to repeat the practice run if they did not feel comfortable with the controls, though none of the subjects felt that a repeat run was necessary. None of the practice images appeared in the main experiment.

For each presentation a single image patch was displayed in the main image window and the subjects were asked to search the image for any clusters of microcalcifications, with a cluster defined as 3 or more microcalcifications. If a cluster was found, the subjects were required to mark the approximate location of its centre by means of the mouse and cursor. Location markers appeared as open white circles 100 pixels in diameter. In the case of large extended clusters that would not fit within the marker, the subjects were asked to just provide a single marker located around the cluster centre, only providing two separate markers when they perceived two separate clusters.

For each marked location, the subjects were asked to give an estimate of their confidence that the marked location actually contained a cluster. Confidence ratings were given on a 5 point scale ranging from "definitely a cluster" (scale point 1) to "probably not a cluster" (scale point 5) and confidence judgements were made by clicking the mouse on the appropriate numbered response button. Each subject was asked to use the whole of the scale and to try to keep their interpretations of the scale points constant throughout the experiment. A sixth scale point "definitely not a cluster" was represented by the "next image" option

that allowed the subject to move on to the next image in the block without making a location judgement.

The subjects were able to make as many location judgements as they wished on an image, each one followed by a confidence judgement. Once they were satisfied the subjects could move on to the next image in the block using the "next image" option.

Prompts, when available, appeared as open red circles 100 pixels in diameter. If an image had prompts associated with it they were displayed automatically when the image first appeared. This initial display lasted for approximately 200 msec before the prompts were removed again. The brief display, combined with the visual prominence of the prompts (red circles on a monochrome background) should have served to alert the subjects to the availability of the prompts, which could then be redisplayed and removed as desired. In addition to alerting the subjects to the availability of the prompts the initial brief display should have served as an attention cue, directing the locus of attention towards the prompted location. However, 200 msec is less than the latency of eye movements so the brief display should not have overly disrupted the search strategy of the radiologist by altering the initial foveal fixation point.

After each block, subjects were asked to give their opinion on how helpful the prompts had been in that block. A questionnaire with a 5 alternative forced choice (5–AFC) format was used to gather this information.

6.3 Results

6.3.1 Methods of Analysis

The true locations of the clusters in the image patches had been determined prior to the experiment by a radiologist working with the original films and pathological

data where appropriate. A location judgement was considered to be true positive if the centre of the marker placed by the subject lay within 50 pixels of the true centre of the cluster. Any location judgement that did not satisfy this criterion was considered to be a false positive.

At each confidence (criterion) level the numbers of true positive and false positive judgements were counted and from these figures the cumulative totals at each level were calculated. The cumulative totals were then converted to probability scores and plotted to generate an ROC curve.

The confidence judgements provided by the subjects show different levels of response criteria, allowing a number of points on an ROC curve to be generated with a single experiment. This is the advantage of rating scale type signal detection experiments when compared to the traditional method that involves running an experiment several times and altering the response criterion of the subject each time by providing different instructions.

In order to perform a statistical comparison of the various experimental conditions, some form of index of detection performance was required in each case. The index used was the signal detection measure A_z, which represents the area under an ROC curve and gives a measure of detection sensitivity. The values of A_z were calculated separately for each of the subjects under each of the experimental conditions and could then be compared by t–tests to determine statistical significance. Ideally an ANOVA should have been used, as it is more powerful than the t–test and shows up interactions between the variables, but unfortunately the cells would have been of different sizes so an ANOVA was not possible.

6.3.2 Analysis of Order Effects

To check for order effects that might have influenced the results, the images in the first and last blocks presented to each subject were compared. Since each

image was repeated (though reversed) in each case there was a set of images that appeared in both blocks and a subset of these that had no prompts associated with either instance. The number of images that fell into this subset varied from 2 to 5 between subjects. A comparison of the first block images with the last block images in this limited group revealed no significant difference between the two (t_{obs} = 1.22). This result suggests that order effects were not a significant problem, though only a very limited amount of data was available for testing and it is not inconceivable that such an effect may have gone undetected.

6.3.3 Sensitivity

Values of A_z were calculated separately for each subject in each experimental condition. These values are shown in figure 6.3. It should be noted that not all of the subjects completed all of the experimental conditions (this is why an ANOVA was not possible). On average it took 20 minutes for a subject to complete one block, so three blocks took around an hour. The demands on the participating radiologists' time meant that an hour was all most of them could spare. In addition, most of the subjects were fairly bored after completing 3 blocks so the extra block would probably have suffered from decreased vigilance due to fatigue. The one subject who managed all 4 blocks completed each block in less than the average time and comparison of the first and last blocks completed (see section 6.3.2) revealed no performance decrease.

Statistical analysis of these results revealed that the detection sensitivity of the radiologists was significantly higher in the level 1 condition (the most accurate prompts) than in the level 2 condition ($t_{obs} = 2.45, p < 0.025$), the level 3 condition ($t_{obs} = 3.79, p < 0.005$) or the unprompted condition ($t_{obs} = 2.28, p < 0.025$). There was no significant difference in detection sensitivity between the level 2 and level 3 conditions ($t_{obs} = 0.33$) and neither of these two conditions showed any

improvement in sensitivity when compared with the unprompted condition (level 2: $t_{obs} = 0.72$, level 3: $t_{obs} = 1.12$).

Subject	Level 1 accuracy	Level 2 accuracy	Level 3 accuracy	Unprompted
1	0.86	—	0.68	0.72
2	0.95	0.79	—	0.92
3	0.90	0.89	0.76	0.77
4	0.91	—	0.89	0.76
5	0.95	—	0.75	0.90
6	—	0.64	0.66	0.86
7	0.93	0.74	0.71	—

Figure 6.3: Summary of A_z values for each subject under each experimental condition.

Values of A_z were also calculated for the prompt generation systems. For the level 1 accuracy system this value was 0.86, which was significantly lower than the detection sensitivity of the radiologists in the level 1 accuracy condition ($t_{obs} = 2.54, p < 0.025$). There was no significant difference between the sensitivity of the level 1 system and the sensitivity of the radiologists in the unprompted condition ($t_{obs} = 1.10$).

The high false positive rates exhibited by the level 2 and level 3 systems made it difficult to calculate any reliable values of A_z for these systems. In both cases the value would fall well below the radiologists' sensitivity levels in both the prompted and unprompted conditions.

If the true positive detection rates are looked at in isolation, it is clear that the radiologists out perform the computer system. Figure 6.4 shows the average TP rates of the radiologists compared to the prompts.

Figure 6.5 shows the ROC curves corresponding to each of the experimental conditions. The curves were generated by averaging the cumulative probability

scores of the seven subjects at each confidence level. The solid lines represent the 3 levels of prompted accuracy and the broken line represents the unprompted condition. Although the graph is fairly untidy it can clearly be seen that the level 1 accuracy condition is set apart from the other conditions which are all intertwined.

	Level 1 accuracy	Level 2 accuracy	Level 3 accuracy	Unprompted
Subjects (Mean)	95%	95%	90%	92%
Prompts	89%	89%	60%	---

Figure 6.4: TP rates of subjects compared to prompts.

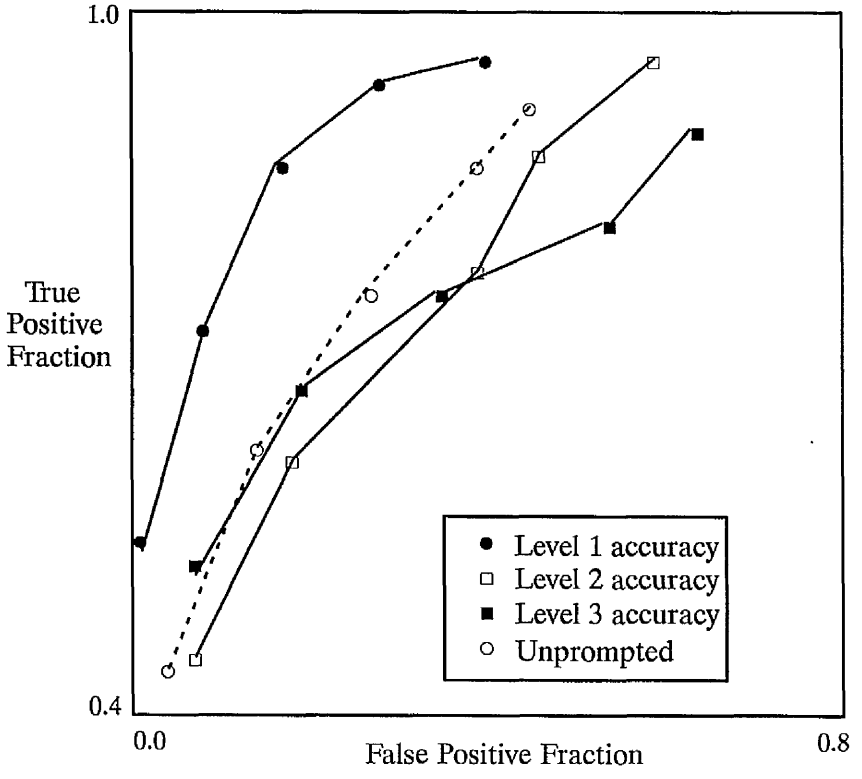

Figure 6.5: ROC curves showing average detection performance in each experimental condition

6.3.4 False Positives

One potential drawback of prompting is that prompts may lead a radiologist to make a false positive judgement when they would not otherwise have done so. This may occur when a false positive prompt directs the radiologist towards a region they might otherwise have disregarded. In this experiment there were 158 false positive judgements made by the radiologists, of which 36 (23%) corresponded to false positive prompts, of which there were 126. Figure 6.6 summarises the number of correspondences that occurred in each experimental condition.

Note that the subjects specialising in mammography (subjects 2, 5 and 6) have much lower levels of correspondence than the general radiologists. It was generally the case throughout the experiment that the mammography specialists demonstrated lower FP rates than the other radiologists.

Subject	Level 1 accuracy	Level 2 accuracy	Level 3 accuracy	Average
1	0	—	5	2.5
2	0	2	—	1.0
3	0	4	7	3.7
4	0	—	5	2.5
5	0	—	2	1.0
6	—	0	2	1.0
7	0	3	4	—

Figure 6.6: Number of radiologists' FP judgements that corresponded to FP prompts

6.3.5 Reading Times

In addition to the accuracy data, the time spent reading each image was also recorded. Figure 6.7 shows the average reading times for each radiologist in each experimental condition.

Subject	Level 1 accuracy	Level 2 accuracy	Level 3 accuracy	Unprompted
1	39.0	—	48.7	21.5
2	6.9	26.6	—	14.5
3	21.0	21.5	21.6	17.6
4	29.1	—	20.2	14.4
5	9.0	—	14.5	9.6
6	—	26.2	26.5	26.1
7	19.3	24.1	23.9	—

Figure 6.7: Average reading times per image (secs).

A statistical analysis of the reading time data revealed that none of the prompted conditions resulted in reading times that were significantly different to the unprompted condition (level 1: $t_{obs} = 0.61$, level 2: $t_{obs} = 1.84$, level 3: $t_{obs} = 1.51$). Similarly, there were no significant differences in reading time between the three prompted conditions (level 1–level 2: $t_{obs} = 0.37$, level 1–level 3: $t_{obs} = 0.63$, level 2–level 3: $t_{obs} = 0.28$).

6.3.6 Subjective Ratings of Helpfulness

Each subject was required to complete 4 item 5–AFC type questionnaire concerning the usefulness of the prompts in each condition. The first three items concerned the helpfulness of the prompts in each block and took the form;

How helpful did you find the prompts to be in block X ?

A Very helpful

B Quite helpful

C Not helpful but not unhelpful

D Quite unhelpful

E Very unhelpful

The fourth item concerned digital mammograms in general and took the form;

How do you feel the reading of digital mammograms compares to original films ?

A Digital mammograms much easier to read

B Digital mammograms generally easier to read

C Little difference between the two

D Original films generally easier to read

E Original films much easier to read

Figure 6.8 summarises the responses of the subjects to these questions. In general the perceived helpfulness of the prompts seems to vary with the accuracy of the prompt generation system. Rather surprisingly the opinions about digital mammography were fairly positive, with three of the subjects rating digital mammograms as easier to read than the original films.

Subject	Level 1 accuracy	Level 2 accuracy	Level 3 accuracy	Digital Mammo.
1	C	—	D	B
2	B	B	—	D
3	B	B	B	B
4	B	—	C	D
5	A	—	B	C
6	—	B	D	B
7	B	C	C	C

Figure 6.8: Subjective ratings of helpfulness of prompts.

6.3.7 Active Use of Prompts

During the experiment, the prompts were presented for an initial brief display and then were available to be switched on and off by the subject. Thus, a distinction may be drawn between the initial passive display of prompts and any subsequent active use initiated by the subject. Figure 6.9 shows the number of times active use was made of the prompts in each experimental condition.

Subject	Level 1 accuracy	Level 2 accuracy	Level 3 accuracy	Average
1	4	—	1	2.5
2	4	22	—	13.0
3	12	17	15	13.7
4	2	—	0	1.0
5	4	—	11	7.5
6	—	1	2	1.5
7	6	11	9	8.7
Average	5.3	12.8	6.3	

Figure 6.9: Number of active uses of prompts in each condition.

A statistical analysis of these data revealed that there were no significant differences in the number of active uses of prompts between the three conditions (level 1–level 2: $t_{obs} = 1.65$, level 1–level 3: $t_{obs} = 0.31$, level 2– level 3: $t_{obs} = 1.19$).

It should be noted that there are extremely large individual differences in the degree to which the subjects made active use of the prompts. It is probably no coincidence that those subjects who made the least active use of the prompts (1, 4 and 6) also gave slightly lower ratings for the helpfulness of the prompts than the other subjects. This suggests that the effectiveness (and certainly the perceived usefulness) of prompting may depend to some extent on the individual concerned and might reflect the differences in the search strategy of individual radiologists.

6.4 Discussion

6.4.1 Interpretation of Results

The results of this experiment support the conclusions of previous studies (Chan 1990, Hutt 1992) that a radiologist working in conjunction with a computer–based detection system is more effective than either the radiologist or the computer system working alone. However, in this case that conclusion only holds true when the prompt generation system performs at a suitably high level of accuracy.

It is interesting to note that the factor which seems to be most important is the false positive rate of the prompting system. If the detection sensitivity data in the level 1 and level 2 conditions are compared, it is clear that the prompts are only useful to the radiologist in the level 1 condition, in which case the FP rate of the prompts is much lower. In both conditions the TP detection rates are equal, suggesting that the high FP rate overrides the moderately high TP rate and eliminates any benefit to the radiologist.

If the benefits of prompting are lost when the prompts fall below a certain level of accuracy, the question is; why does this happen ? One possibility is that, as often happens in many cases of decision making, the radiologist is performing a form of cost/benefit analysis on the prompts in order to determine whether or not they are worth accounting for in their search strategy. In a situation such as the level 1 condition, when the prompts are predominantly true positive, the benefits of attending to all of the prompts outweigh the slight costs of pursuing 'red herrings' when the prompts are invalid (false positive). Consequently, the radiologist accepts the usefulness of the prompts, accounts for them in their search behaviour and there is an effect on detection performance. However, as the false positive rate increases, so too do the costs of attending to the prompts, in terms of the extra workload in checking false positives as well as the increased disruption of the normal search strategy. At some point the costs begin to exceed the benefits gained from attending to the prompts and the radiologist ceases to allow the prompts to affect their search patterns; in short, the prompts are ignored.

If this explanation is true then we would not expect to see the prompts in the level 2 and 3 conditions influence the detection performance of the radiologists. The analysis of detection sensitivity shows that in these two conditions detection performance is not significantly different from the unprompted condition, suggesting that the prompts in these conditions are having no effect. Similarly, the reading time data suggest that the average reading time is not affected by the

presence of prompts in the level 2 and 3 conditions compared to the unprompted condition. The fact that there is a substantial amount of extra information in the prompted conditions, but no apparent increase in reading time, suggests that the extra information is not being processed by the radiologist. Of course, it might be the case that the time required to process and act on the extra information is negligible and therefore does not show up in the figures, but a previous study, (Hutt 1992), has shown that prompting can lead to increases in reading time, suggesting that this is not the case. One major problem here, is that the level 1 condition does not show any significant differences in reading time compared to the unprompted and other conditions, even though there is a clear effect of the prompts on detection sensitivity. One possible explanation is that the effectiveness of the prompts in rapidly directing the attention of the radiologist towards the clusters, leads to a reduction in search time that counteracts the extra time required to process the prompt information.

The cost/benefit analysis explanation assumes that at some point the radiologist makes a decision that the prompts are no longer worth the increased workload and should be ignored. It would therefore be expected that the radiologist's perception of the helpfulness of the prompts would decrease in the lower accuracy conditions and this is observed to be the case. Five of the seven subjects gave a lower rating of helpfulness to the prompts in level 2 and/or level 3 conditions than in the level 1 condition, while the ratings of the other two subjects remained constant throughout. No subjects gave a higher rating of helpfulness to any other condition than they did to the level 1 condition.

However, there is some evidence to suggest that the subjects were not ignoring the prompts in the level 2 and 3 conditions as proposed by the cost/benefit analysis explanation. The main problem lies in the data concerning the active use of prompts. An analysis of the active prompt use data shown in figure 6.9 showed that there was no significant difference between the three prompted conditions

in the number of times prompts were actively used. If the prompts were being ignored, we would expect to see little or no active use of the prompts, and certainly we would expect significantly less active use than in the level 1 condition in which prompting clearly has an effect on detection sensitivity. While it seems as though the prompts were used to the same degree in all the conditions, it should be noted that the data in figure 6.9 refer to the *number* of times the prompts were actively used, which does not account for the number of prompts that were available in each condition. A previous study, (Hutt 1992), has suggested that prompts are actively displayed more often when there are multiple prompts associated with an image, while in this experiment the level of active prompt use seems to remain constant regardless of the fact that the number of images with multiple associated prompts varied greatly. In the level 1 condition only 37.5% (6 out of 16) of the prompted images had more than one prompt, while in both of the level 2 and level 3 conditions the proportion of images with multiple prompts was 87.5% (21 out of 24). In addition, none of the images in the level 1 condition had more than 2 associated prompts while 14 of the prompted images in the level 2 condition and 13 of the prompted images in the level 3 condition had 3 or more prompts associated with them. As a consequence of all this we might expect to see substantially more active use of prompts in the level 2 and 3 conditions if the degree of prompt use across all three prompted conditions was equivalent. Therefore, although the number of times that prompts were used may be effectively equivalent across the three prompted conditions, it may be the case that relative to the number of images with multiple prompts in each condition, the level of active prompt use in the level 2 and 3 conditions is actually less than in the level 1 condition.

Of course, this analysis still does not explain the observation that there is a significant level of active prompt use in level 2 and 3 conditions when it has been suggested that the subjects are ignoring the prompts. One possible explanation of this inconsistency concerns the way in which the prompts are being used. So

far we have only considered the prompts as attention cues – influencing the search pattern of the subject. However, prompts may have a secondary function as reinforcers or 'second opinions'. It is possible that the subjects are ignoring the prompts when it comes to altering their search strategy, but having located a potential cluster that they are not sure about, the subjects actively display the prompts to see if the computer agrees with them. This would explain why the level of active prompt use does not seem to be affected by the number of multiple prompts; the subjects are not recalling the prompts to check each prompted location, but only to see if one of the prompts corresponds to the location they have identified as suspicious. If this is the case, the active use does not depend on the number of multiple prompts, but rather on the number of occasions that the subject requires a second opinion.

Further evidence against the cost/benefit analysis explanation comes from the data concerning the correspondence between false positive prompts and false positive judgements made by the radiologist. These data show that there were no cases of correspondence in the level 1 condition, but a fair number in each of the other two prompted conditions, especially the level 3 condition. This seems to suggest that the large numbers of false positive prompts in the level 2 and 3 conditions are leading the radiologist to make false positive judgements when they might not otherwise do so. It should be noted that images containing several prompts, as are common in the level 2 and 3 conditions, have a substantial proportion of their area associated with prompted regions, so the possibility of a radiologist's false positive judgement and an invalid prompt coinciding purely by chance are increased. Nevertheless, pure coincidence is not enough to fully account for all of the correspondence observed – it seems as though the relatively high levels of false positive prompts do increase the false positive rates of the radiologists.

The above observation does rather contradict the idea that the prompts are being ignored when the false positive rates are high, since if that were the case it does not seem possible that the invalid prompts could be leading the radiologists to make false positive judgements. However, the correspondence of false positives could fit the idea of prompts being used only second opinions, since the false positive prompts may then act to reinforce the radiologists opinion that an abnormality is present when it is not the case.

The correspondence between false–positive prompts and false–positive judgements raises another question; If the false positive prompts in the level 2 and 3 conditions are leading to increases in the radiologists' false positive rates, why are the observed detection sensitivities in these conditions no worse than that in the unprompted condition ? One possible answer is that prompting does actually lead to some improvement in the detection rate of the radiologists even in the low accuracy conditions, possibly by acting as a second opinion. However, this improvement may be offset by the increase in the radiologists' false positive rates brought about by the high numbers of invalid prompts, so that the two factors cancel out – leading to no effective improvement in the overall detection sensitivity.

6.4.2 Limitations of the Experiment

An important feature of any experiment is the extent to which the results may be generalised to other situations and though this experiment has generated some interesting data concerning the ways in which errors in the prompt generation system affect the performance of radiologists, there are a number of factors that bring into question the extent to which the results may be generalised.

Firstly, the low number of subjects, coupled with the limited size of the data set, makes it very difficult to be sure of the reliability of the results. An additional problem occurs because of the need to use radiologists who are not particularly

experienced in mammography to make up the numbers. Eliminating these subjects from the study leaves such a small subject group that no worthwhile data could be generated.

The experiment also suffers from the highly artificial nature of the task. This is perhaps the most significant problem that prevents the generalisation of the results to any form of non-experimental setting. Only partial views of a single mammogram were used in the study, which is far removed from the normal film reading situation when films from both the left and right breasts are viewed simultaneously. In addition, the presentation of the image on a computer screen is very different from the usual display format of film and light box.

Additional problems occur in the area of the constraints on the subject's responses. The confidence rating scale is a useful technique that allows a number of points on an ROC curve to be generated in one study, but it is very difficult to be sure that each subject's precise interpretation of the scale points has remained constant throughout, and even more difficult to be sure that all of the subjects have used the same interpretation. Furthermore, the subjects were generally used to being much freer in the range of possible diagnoses that are available to them. Some of the subjects seemed to lose sight of the fact that *clusters* of *micro*calcifications were the only abnormalities they were supposed to be marking.

6.5 Conclusions

The experiment has served to further confirm the results of earlier studies that a radiologist working in conjunction with a prompting system is more effective than either the radiologist or the system working alone. It has also demonstrated that the beneficial effects of prompting are lost if the accuracy of the prompts falls too low, particularly when the rate of false positive prompts increases.

The lack of improvement in the radiologists' detection sensitivity that occurs with a high false positive rate may be a consequence of the large amount of invalid information leading the radiologist to disregard the prompts when searching the image. Alternatively, there may be some improvement in the true positive detection rate of the radiologists even when the numbers of invalid prompts are relatively high, but this improvement may be counteracted by an increase in the radiologists' false positive rate caused by the invalid prompts, so that there is no apparent overall improvement in sensitivity.

Although the data generated by the experiment are quite interesting, it would be very difficult to relate the results to anything approaching a clinical environment due to the highly artificial nature of the task and the limited numbers of expert subjects who took part. Consequently, in order to establish some more meaningful results, an experiment would need to simulate the task of mammographic film reading much more closely.

Chapter 7

Prompting in a Realistic Environment

7.1 Objectives

The results of the study described in chapter 6 are encouraging in that they show that under certain circumstances, prompting can be a useful aid to the detection of Breast Cancer. However, it is difficult to generalise the results of this study to a clinical setting due to the artificial nature of the experiment. The main aim of this second study is to examine the effects of prompting in as realistic an environment as possible, while retaining sufficient experimental control to ensure that meaningful results are obtained.

The underlying philosophy in the design of this second experimental study was to mimic, as closely as possible, the film reading task as it would be undertaken in a typical screening environment. To this end, the experiments took place in actual screening centres using the film viewing equipment employed by radiologists during a routine screening session.

Apart from investigating prompting in a realistic setting, there are three main aims to this study. Firstly, the experiment will examine whether prompting is an effective aid to the detection of well–defined lesions. At present all of the studies

that support prompting as a CAD technique have relied on studies of prompted microcalcifications. In this experiment, computer-based systems will be used to generate prompts both for microcalcifications and for well-defined lesions, making it possible to compare the effectiveness of prompting these two classes of abnormality.

A second aim of this experiment is to examine the effects of prompting when multiple abnormalities are present in the image. In the past, prompting studies have concentrated on images containing only a single abnormality, but it may be the case that more than one abnormality is present in a mammographic study such as in the case of bilateral breast cancer, for example.

The third aim is to investigate the effects of prompting on untargeted abnormalities. In this experiment prompt generation systems will be used to target clustered microcalcifications and well-defined lesions, but images in the experiment will also include examples of other abnormalities that are not targeted by the prompting systems, such as spiculated lesions and architectural distortion. The inclusion of images with a wide variety of mammographic abnormalities, rather than presenting only a single type, also serves to bring the study closer to the normal screening task.

7.2 Experimental Method

7.2.1 Images

In total, 100 pairs of mammographic films were used in the study. Each pair consisted of the mediolateral views of the left and right breasts of a single patient. All of the films were produced from routine screening. Each mammogram pair belonged to one of six experimental groups as follows:

- Group 0: Normals; no abnormality of any type present in either mammogram of the pair *(50 pairs)*.

- Group 1: Clusters; a single cluster of microcalcifications present in one of the mammograms in the pair *(10 pairs)*.

- Group 2: Lesions; a single well–defined lesion present in one of the mammograms in the pair *(10 pairs)*.

- Group 3: Untargeted; a single untargeted abnormality in one of the mammograms in the pair. In six cases this was a spiculated lesion and in four cases this is an architectural distortion *(10 pairs)*.

- Group 4: Multiple targeted; two or more abnormalities that may be clustered microcalcifications or well–defined lesions or a combination of the two. Multiple abnormalities may appear in a single mammogram or in both *(10 pairs)*.

- Group 5: Multiple untargeted; one or more abnormalities that may be clustered microcalcifications or well–defined lesions or a combination of the two plus one or more untargeted abnormalities. Multiple abnormalities may appear in a single mammogram or in both *(10 pairs)*.

Within groups 1–5 above, half of the films contained malignancies and half contained only benign abnormalities. In groups 4 and 5, when multiple abnormalities were present, some of the cases containing malignancies also contained benign abnormalities.

In this data set there was a normal to abnormal image ratio of 1:1, though in clinical practice only around 10% of screening films will show any signs of abnormality. In keeping with the concept of maintaining realism throughout the experiment, it might be appropriate to have the normal to abnormal ratio similar to that found in practice. However, in order to maintain this ratio, 500 normal pairs would also be required, each of which would be need to be presented both

with and without prompts for a total of 1100 presentations. This is rather more than the participating radiologists could be expected to read for an experiment.

Each of the films was digitised with a sampling rate of 100 μm per pixel and each digital image was then divided into five slightly overlapping 1024x1024 image patches in order to be processed by the prompt generation algorithms. Each patch was processed by the fuzzy pyramid algorithm, as described in section 5.2, both for the purposes of locating microcalcifications and for the detection of well–defined lesions. For the detection of microcalcifications an *upper* link strength threshold of 0.90 was used, while for detecting lesions a *lower* link strength threshold of 0.55 was used.

It was observed that, after processing with the prompt generation algorithm for the detection of microcalcifications, approximately one false–positive prompt per image patch was generated. This false–positive rate compares favourably with many of the systems that have been described for detecting microcalcifications (see section 4.2). However, since there are five such patches in one full image and two films to the pair, this generated an average of around 10 false–positives per image pair – a clearly unacceptable number of false–positives – making it necessary to remove some of the invalid prompts (effectively simulating a prompt generation system of much greater accuracy). This was achieved by performing the analysis again with a much lower threshold; 0.30, which generated very few false–positives or true–positives. the false–positive prompts from this procedure were then combined with true–positive prompts from the original processing in order to produce the final result. A similar procedure was used for the lesion detection system with a second threshold of 0.85.

Of all the microcalcification clusters used in the experiment, 86% were prompted, while of all well–defined lesions, 65% were prompted. Although the prompt generation algorithm was not designed to detect the untargeted abnormalities, some of the untargeted abnormalities were actually found by the prompting

system and as a consequence 35% of the untargeted abnormalities were also prompted. The false–positive rate across all of the image pairs in the experiment was 1.1 false–positives per pair. Figure 7.1 shows the prompt accuracy data broken down by image group.

Image Group	True–positive rate	False–positive rate
Group 0: Normals	---	1.0 pair^{-1}
Group 1: Clusters	90%	1.2 pair^{-1}
Group 2: Lesions	70%	1.1 pair^{-1}
Group 3: Untargeted	40%	1.2 pair^{-1}
Group 4: Multiple targeted	64%	0.9 pair^{-1}
Group 5: Multiple untargeted	57%	1.0 pair^{-1}

Figure 7.1: Prompt accuracy according to image group

Each of the image pairs was printed out on a laser printer to produce two low resolution hard copies, each of which showed a pair of mammograms laid out as they might be presented to the radiologist on a film viewer. Each hard copy also contained the film number and rating scale for the radiologist's response.

Of the two hard copies of each film, one showed the film pair in its unprocessed form with no prompts, while the other showed the processed version with the prompts presented as open white circles superimposed on the digital mammograms.

Appendix 1a shows both hard copies for one film pair containing a single microcalcification cluster. The prompted version contains one true–positive and two false–positive prompts.

The low resolution hard copies all contained sufficient detail to be recognised as the corresponding original films but only in very few cases is enough detail present for the abnormalities to be visible.

7.2.2 Subjects

Eight practising mammographic radiologists participated in the study. All of the participants were either consultant radiologists or senior registrars and all were involved in regular mammographic screening at various sites in the UK.

7.2.3 Procedure

Each radiologist was presented with each image in the experiment in both the processed (prompted) and unprocessed (control) conditions. Since there were 100 film pairs, this entailed 200 presentations which were given in two sessions of 100 presentations each. Gale (1989) has demonstrated that radiologists are able to read 100 film pairs in an hour of continuous reporting with no reduction in performance due to lack of vigilance. In addition, consultation with radiologists has revealed that a typical screening session might include 100 or more film pairs and would be expected to last approximately an hour. It would seem, therefore, that 100 films in a session is a realistic number to expect the subject radiologists to read and a vigilance decrement is not expected to affect the results.

Since each subject was reading each film twice, once with and once without prompts, the sessions were separated by a number of days and no two presentations of the same film occurred in the same session. This should have helped to prevent learning from affecting the results by reducing the chances that the radiologists will remember the films. Gale (1979) has suggested that it is not uncommon for radiologists to give different diagnoses from the same film on different occasions, so if a film is remembered from the previous session, it is quite possible that the response will not be. Even if the response and film were remembered, the radiologist had received no feedback concerning the correctness of their previous response, so remembering the film should not have been any help to them in making a judgement. In addition, each of the radiologists was involved in regular mammographic screening and all of them would have been presented with a large number of mammograms between the two sessions.

In each case the radiologist was presented with the original pair of mammograms on a film viewer in their screening centre. In addition to the original films, one of the hardcopies was also presented; either the processed or unprocessed version depending on the experimental condition. The films were loaded on to the viewer in the same way they would have been during normal film reading with all of the films for that session loaded on to the viewer in two blocks – fifty of one experimental condition (either prompted or control) followed by fifty of the other condition. The hard copies were presented in a similar way to that used for the patients' medical records in a normal screening session; they were stacked to one side in the same order as the films appear on the viewer.

The hardcopies served as media for the prompt information when such information was available, but also acted as response forms. Each hardcopy displayed a film number for reference, the copy mammograms (with or without prompts superimposed) and a six point rating scale of the following form;

0: Normal

1: Benign

2: Probably Benign

3: Uncertain

4: Probably Malignant

5: Malignant

The points for the rating scale correspond to those generally used to rate films in screening centres with the exception of the '0: Normal' point. Common practice in screening is not to give any rating for definite normals since these films are simply archived and no further action is taken.

Each radiologist was asked to study the original pair of mammograms on the viewer and provide a rating for it by ringing the corresponding point of the rating scale on the hard copy associated with that film pair. In every case where the

radiologist gave a rating above zero, they were requested to mark the locations of any and all abnormalities on the hard copy of the image pair before moving on to the next pair.

When prompts were presented they should have been used in conjunction with the original films to locate potentially suspicious regions. Prior to the experiment, the radiologists were told the approximate accuracy levels of the prompting system and the fact that it was only targeting microcalcifications and well defined lesions, so that they knew what to expect rather than alter their strategy after a few films because they have made their own decisions about accuracy levels. They were also informed of the ratio of normal to abnormal films for similar reasons.

The radiologists were not given any practice runs before the main session as the task should have been familiar enough to make it unnecessary. However, in order to explain what was expected of them, they were shown a few examples of hard copies both with and without prompts when the experimental task was explained to them. None of the examples included films that appeared in the experiment.

For each film in the study a transparent overlay was produced with the edges of the hard copy images and the locations of any abnormalities marked on it. These known locations were provided by a consultant radiologist who had viewed all of the original films in conjunction with the patients' medical records. The overlays were used to check the responses of each subject after the experiment and calculate the numbers of true– and false–positives at each point on the rating scale. This provided data suitable for ROC analysis and subsequent statistical comparison of the experimental conditions.

7.3 Results and Discussion

The detection performance of each radiologist under both the prompted and control conditions was determined by means of ROC analysis. This involved

calculating the number of true–positive and false–positive responses at each point of the rating scale, since these points represent different levels of response bias. These values were then plotted to yield a free response operating characteristic (FROC) curve for each of the subjects. The individual FROC curves were then pooled to yield the curves shown in figure 7.2 by averaging the numbers of true–positives and false–positives at each criterion level.

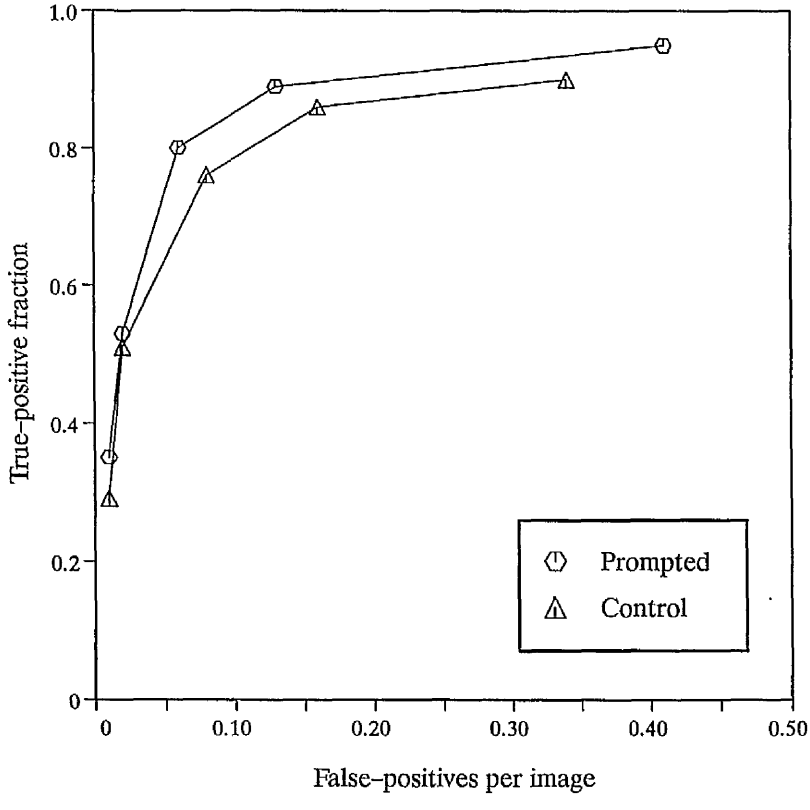

Figure 7.2: Pooled FROC curves showing detection performance of radiologists in each experimental condition

In order to perform conventional ROC analysis, the fractions of true–positives and false–positives at each level of response bias were required. These were obtained by collapsing the 'normal' and 'benign' scale points into a single category and taking the highest rating for each film as the criterion level. Once again, an

ROC curve was prepared for each participating radiologist and the results were pooled by averaging at each level of response bias. The pooled ROC curves for each experimental condition are shown as dotted lines in figure 7.3, with the solid lines representing the best-fit ROC curve determined by maximum likelihood estimation (Metz, 1989).

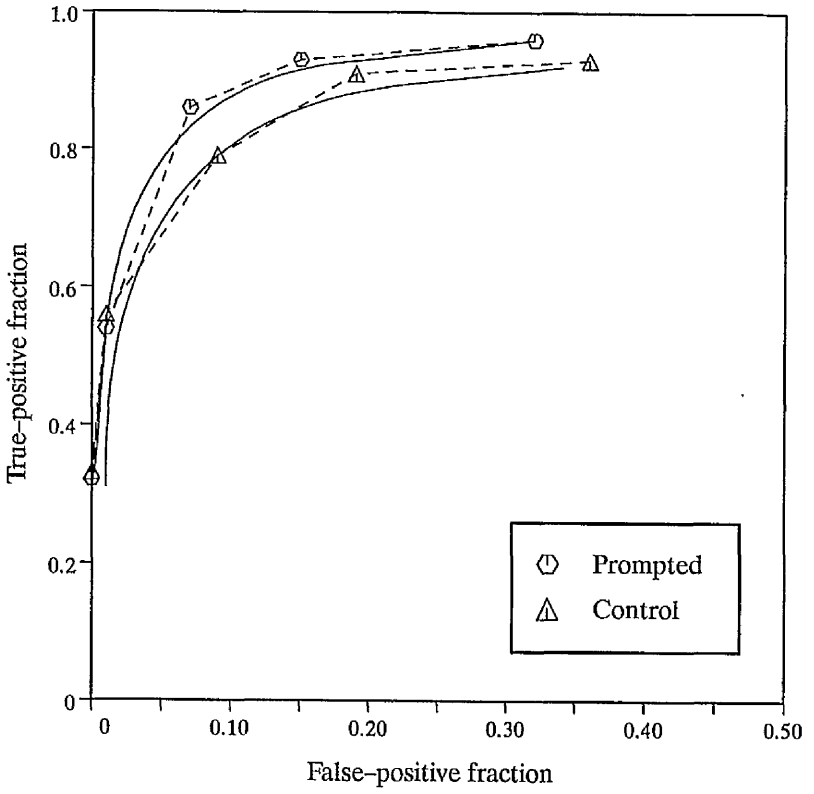

Figure 7.3: Pooled ROC curves showing detection performance of radiologists in each experimental condition

In order to test for statistical significance, values for the signal detection measure d_a were calculated for each radiologist in each of the two conditions. These values are summarised in figure 7.4. An analysis of these values showed that the detection performance of the radiologists in the prompted condition was significantly better than their performance in the control condition ($t_{obs} = 4.13$, $p < 0.005$). It is

interesting to note that the general case is also true for each individual radiologist with all of the participants demonstrating an improved level of performance in the prompted condition.

Subject	Prompted	Control
1	2.85	2.32
2	2.53	1.98
3	2.66	2.31
4	1.89	1.77
5	2.76	2.62
6	3.06	2.33
7	2.91	2.61
8	2.63	1.97

Figure 7.4: Values of d_a for each subject

The detection performance of all the participating radiologists was substantially higher than that of the prompt generation algorithm, which had an average true–positive detection rate of 51 out of 75 (68%) at a false–positive rate of about 1.1 invalid prompts per image. These results confirm Chan's suggestion that a radiologist working in conjunction with a computer–aided diagnostic system is more accurate than either the radiologist or the system working alone.

The results of the previous experiment suggested that false–positive prompts might reduce the effectiveness of a prompting system by leading radiologists to make false–positive judgements that they might not otherwise have made. Figure 7.5 shows the total number of false–positive judgements made by each subject in each of the two experimental conditions in the second set of experiments. There is a tendency for there to be more false–positives in the prompted condition than in the control, though this difference is not statistically significant ($t_{obs} = 1.36$).

However, in this case there were relatively few invalid prompts in the images. It may be that the detrimental effect of invalid prompts on the detection performance of radiologists is relatively small and requires a large number of invalid prompts before it becomes significant.

Subject	Prompted	Control
1	56	47
2	13	17
3	68	39
4	27	28
5	34	29
6	19	18
7	31	27
8	22	14

Figure 7.5: Number of false-positive responses in each condition

It is clear that the reason for the increased detection performance in the prompted condition is that the radiologists correctly detected prompted abnormalities that were missed in the control condition. A total of 600 abnormalities were presented in the experiment (75 to each of 8 radiologists) and 37 of these were only identified in the prompted condition. However, there were 4 abnormalities that were missed in the prompted condition but correctly located in the control condition. These 3 abnormalities all had one things in common; they were all missed by the prompt generation system. In addition, in three of these four cases there was at least one invalid prompt present in the image. It is possible that this combination of a false-negative and false-positive error on the same film could result in the prompt having the opposite effect to that intended, directing attention away from an abnormality towards a region of normal tissue and thus causing the radiologist to miss the abnormality. This could be a serious problem for a prompting system.

However, there were 16 cases of this type of combination error among the 100 film pairs used in the study, giving a total of 128 instances across the eight subjects. In only four of these instances was a problem observed. The problem seems to be fairly minor in comparison to the advantages of prompting, though it should not be ruled out completely. A double–reading system where one radiologist used prompted films and one did not should eliminate any potential missed abnormalities caused by combination errors,

In addition to examining the radiologists' performance as a whole, the results were broken down among the various types of image in the study. Table 7.6 shows the pooled values of d_a for each of the types of image under each experimental condition.

Type	Prompted	Control
Single Cluster	2.81	2.29
Single Tumour	2.68	1.98
Single Other	2.27	2.31
Multiple targeted	2.61	2.18
Multiple untargeted	2.41	2.35

Figure 7.6: Values of d_a for each type of image

An analysis of the d_a values for the various types of image revealed that the detection of clustered microcalcifications was improved in the prompted condition ($t_{obs}=3.03$, $p<0.025$) as was the detection of tumour masses ($t_{obs}=3.85, p<0.01$) and the detection of lesions in film pairs containing multiple examples of clusters and tumours ($t_{obs}=2.66, p<0.025$). There was no significant difference between the two conditions for the single untargeted abnormalities group ($t_{obs}=0.73$), but since few of the abnormalities in this group were actually prompted, the prompted versions of the films in this group were similar to the

control versions. There was no statistically significant improvement in performance with prompting for films containing multiple abnormalities with at least one untargeted lesion, though there was a slight, non-significant increase. Although a number of target abnormalities in this group were prompted, the majority of the untargeted abnormalities were not, reducing the effect of prompting for the images in this group.

7.4 Summary & Conclusions

Although the number of radiologists participating in this study was fairly low, the experiment was designed to capture some of the important aspects of screening in a realistic environment and consequently the results can be much more reliably generalised to the way in which prompting might work in clinical practice.

This is particularly encouraging, as once again the benefit of prompting to the radiologist has been demonstrated, certainly in the case of single clusters and lesions, and even with multiple targeted lesions on the the same film pair. It is also encouraging that the prompts did not seem to act as distractors in the case of the untargeted abnormalities – as detection rates for these was no lower than in the unprompted condition.

One slight concern is the issue of combination errors and whether these might serve to reduce detection performance in certain cases by directing attention away from the locations of potential abnormalities. Once again, the role of false-positive prompts seems to be important. False-positives provide no valuable information and in certain cases may act as distractors – to the detriment of film reading performance.

The next study (chapter 8) will examine the effects of false-positive prompts in more detail and will suggest a possible model to explain why these invalid cues may act as distractors in certain cases but not in others.

Chapter 8

The Relationship between True–positive and False–positive prompts

8.1 Introduction

8.1.1 A Possible Relationship

Previous experiments on the effectiveness of prompting have suggested that an excessive level of invalid (false–positive) prompts can reduce or even eliminate the benefits of a prompting system (see chapter 6). This suggests that in order to develop an effective prompting system it is important to establish what constitutes an acceptable level of error in prompt generation.

There appear to be two main reasons why high levels of false–positives may reduce the effectiveness of prompting. Firstly, a false–positive prompt may lead a radiologist to make a false–positive judgement that would not otherwise have been made – counteracting any increase in true–positive detections due to prompting. This is an unfortunate side–effect of an automatic prompt generation system that operates by searching for structures with characteristics typical of mammographic abnormalities. Any structures flagged by such a system may well

have an appearance similar to malignancies. If the radiologist considers the detection system to be effective in detecting abnormalities, the invalid prompt may act to reinforce the opinion that a normal structure is actually abnormal. This effect would be more likely to occur with less experienced radiologists.

In cases such as these it is very difficult to set a general value for an acceptable false–positive rate. Any such value could be expected to depend heavily on the algorithms used by a particular prompt generation system. If the algorithms rely heavily on search strategies similar to those employed by human observers, false–positive errors may often lead to the prompting of structures that have an appearance similar to abnormalities. The problem is likely to be less severe when the detection system uses different techniques to those employed by radiologists – as it is then possible that any resulting false–positives will look less like genuine abnormalities.

This problem could be circumvented by the use of double–reading. If only one reader were to be provided with prompts, consultation between readers should serve to weed out any false–positive judgements produced solely as the result of prompting.

The second reason why high levels of invalid prompts may reduce the effectiveness of the system is simply that the radiologist is wasting time checking a large number of obviously benign structures. If many such structures have been prompted then the system may be perceived as being too inaccurate and the prompts may be ignored. In this case a lot of time and effort has been expended developing a prompting system that is not being used.

In the study described in chapter 6 the prompts were only of benefit to the user in one experimental condition. In this condition, the true–positive rate was high (approx 90%) and since half of the 96 images contained abnormalities, there were 43 true–positive prompts in the data set. The false–positive rate in this condition was relatively low (approx 0.5 image^{-1}) so there were 47 invalid prompts in the

data set. In other words the numbers of true– and false–positive prompts in this condition were roughly equal, so that any given prompt had an almost equal probability of being valid or invalid. In the other experimental conditions, the numbers of invalid prompts greatly exceeded the numbers of true–positives, so any given prompt was far more likely to be invalid than valid. In these conditions there was no benefit from prompting.

This result may suggest a method of setting an acceptable false–positive rate so that a system is perceived as being sufficiently accurate to be of use. It is possible that as long as any given prompt is at least as likely to be valid as invalid then the false–positive prompts may not lead to a reduction in the effectiveness of the prompting system.

8.1.2 Objectives

The results from the experiment cited above, although suggestive, are far from being conclusive. An experimental investigation is required to test whether these ideas have any merit.

The aim of this study is to test the hypothesis that prompting is effective only if the number of false–positive prompts does not greatly exceed the number of true positives. If this is the case then it may be useful in the development of prompting systems, as it indicates a way in which an acceptable false–positive rate may be set.

8.2 Experimental method

The ideal way to test this hypothesis would be to present a large quantity of mammograms to a number of radiologists. Each mammogram would be presented with prompts that are generated with one of several different

combinations of true– and false–positive rates. In order to prevent learning from affecting the result, any given radiologist should see only one combination of true–positive and false–positive rates. This would require an impracticably large number of radiologists to participate in the study.

In order to conduct a large–scale realistic study, it is necessary to identify a critical range of true–positive/false–positive rates, so that the number of conditions required in the full–scale study could be kept to a minimum. To this end, a smaller scale preliminary study was carried out with a wide range of prompting accuracy rates.

For the purposes of this preliminary study a *simulated* mammogram reading task was used. This task was designed to encompass as many elements of the true film–reading task as possible but did not employ mammograms and did not therefore require radiologists to read them. This allowed a large number of non–radiologist subjects to participate in the experiment so that a wider range of true–positive/false–positive rate combinations could be studied.

8.2.1 Elements of the Mammogram Reading Task

Mammographic film reading can be considered as a signal detection task in which the observer must locate one of several target types that may vary in appearance and may or may not be present within a background of structured noise. In order for the results of the experiment to be of any value, the simulated task should reflect as many of the important elements of genuine mammogram reading as possible without actually employing mammograms. It is therefore necessary to identify the important aspects of the film–reading task that should be simulated:

- **Target appearance.** There are many classes of mammographic abnormality with a high degree of inter– and intra–class variability. Abnormalities may also vary in size, contrast, definition and orientation.

- **Target familiarity.** Although there is a certain amount of target variation, radiologists are familiar with the typical characteristics of abnormalities.

- **Background.** Normal breast tissue can also vary greatly in appearance depending on the level of fatty/glandular tissue present. A number of normal structures (ducts, vessels etc.) may also be present.

- **Anatomy.** A radiologist is able to ignore certain structures in the mammogram, since they have a level of expertise in the anatomy of the breast that is used to guide their search strategy. This would be extremely difficult to simulate.

- **Quantity/Time.** A typical screening session may involve around 100 patients (200 films). In some cases a film may only be scanned for a few seconds. This scanning time is not fixed and the radiologist is free to examine any film further.

8.2.2 Stimulus Images

Each subject participating in the experiment was presented with a series of images. These images represented the simulated mammograms, though no attempt was made to mimic the appearance of a mammogram, just to retain the important elements of the task. A total of 100 images were used, 50% of the which contained some form of target. The targets present in the images varied in appearance and were selected to reflect the appearance of mammographic abnormalities, as illustrated in figure 8.1.

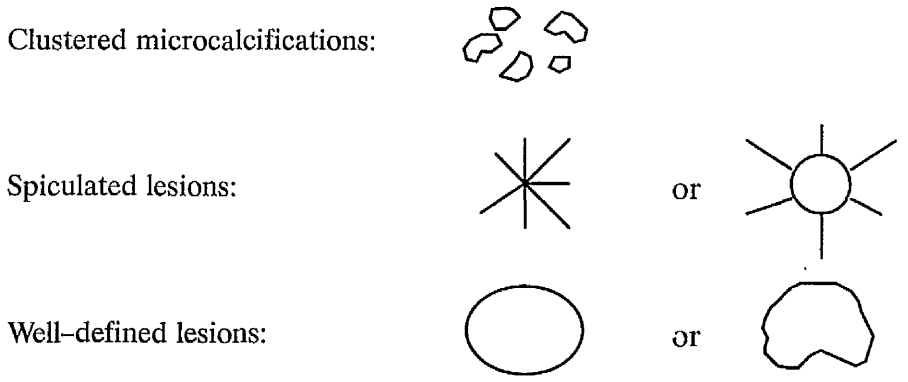

Figure 8.1: Examples of typical appearance of targets.

Targets were presented with various orientations and sizes. No two targets in the data set were identical. Other forms of variation specific to individual target type were also used, ie. the number of 'microcalcifications' in a cluster, the number of 'spicules' associated with a lesion and the shape of well–defined lesions. The participating subjects were familiarised with the general appearance of the targets during a brief instruction and training period before the main experiment.

Each target was presented within a field of structured noise. In order to prevent any ceiling or floor effects in the results, the task was designed to be difficult enough to prevent any subject from scoring 100% but simple enough to prevent subjects from scoring zero. Treisman's work on pre–attentive processing suggests that targets will be more difficult to detect if they are embedded within a field of distractors that have features in common with the targets. These distractors included; straight and curved lines, individual 'microcalcifications', geometric objects and regions of varying colour.

Figure 8.2 shows an example of one of the images used in the experiment. In this example the density of distractors is moderately high and since the precise form

of the target is unknown, detection is fairly difficult. The task can be made either easier or harder by altering the density of distractors. The experiment used images with various distractor densities. The target in figure 8.2 is a simulated spiculated lesion in the upper–left region of the image.

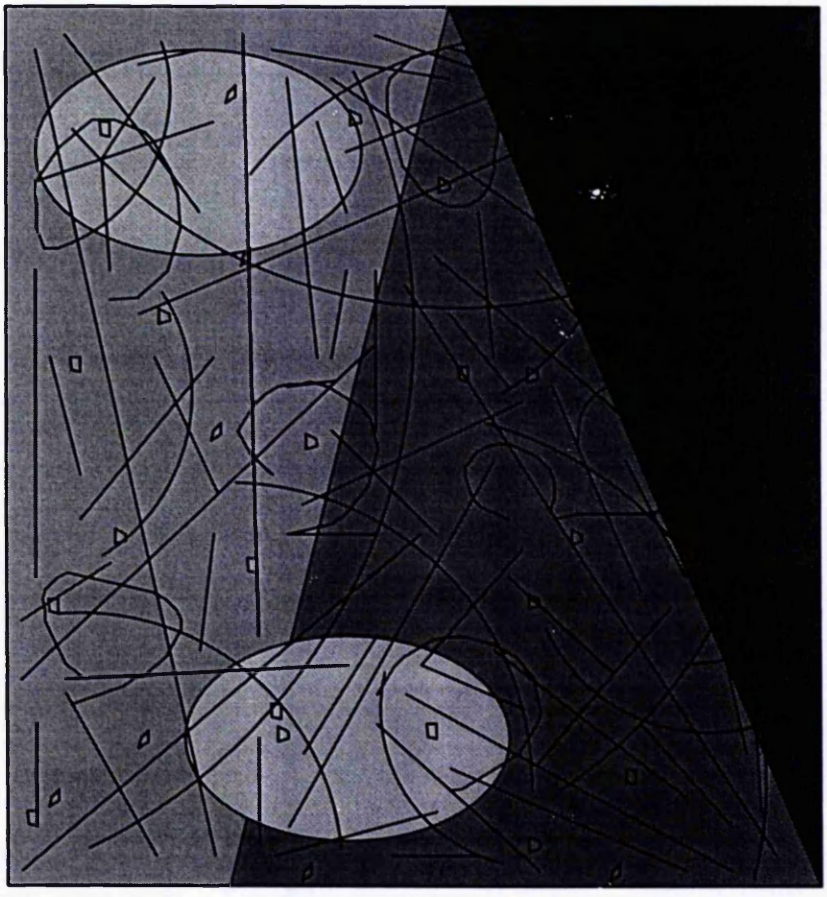

Figure 8.2: Example of Image used in Study

The image backgrounds were generated randomly. The backgrounds consisted of patches of varying intensity with a number of distractors superimposed. These distractors varied in size and shape and generally had characteristics in common with the targets. Half of the randomly generated backgrounds were randomly selected and targets were added to these. The targets also varied in appearance – the aim being to produce a range of images – some of which contain relatively

easy to detect targets and some in which targets are very difficult to detect. The lack of an exact specification for the appearance of a target should have served to make the task more difficult.

8.2.3 Procedure

A total of 90 subjects took part in the experiment, each of whom was assigned to one of 9 experimental conditions. These conditions represented different combinations of true–positive and false–positive error rates.

The 9 experimental conditions were derived by combining three different true–positive rates (100%, 75% and 50%) with three different false–positive rates (0.17 im^{-1}, 0.33 im^{-1} and 0.5 im^{-1}). Figure 8.3 shows the 9 conditions and the ratio of the number of true–positives to the number of false–positives in each condition.

	Number of False–positive Prompts			
Number of True–positive Prompts		**8**	**16**	**24**
	16	0.5	1.0	1.5
	12	0.66	1.33	2.0
	8	1.0	2.0	3.0

Figure 8.3: Ratio of true–positive to false–positive prompts in each experimental condition

Each of the 90 subjects was presented with 96 images, 48 prompted and 48 unprompted – reflecting the number of presentations that might typically take place in a screening session. Each image was presented for 6 seconds in order to

introduce time pressure and increase the difficulty level of the task, thus preventing ceiling effects. The images were presented on A4 sheets bound together in a booklet.

After each presentation, the subject was required to make a simple Yes/No judgement on whether a target had been present. These responses were recorded on a separate score sheet.

One in three of the images contained targets in both the prompted and unprompted conditions. In no case did more than one target appear in a given image. The three types of target illustrated in figure 8.1 were presented in equal numbers.

Prior to the main experiment, the subjects received a training session in which the task was explained by means of both written and verbal instructions. Examples of typical targets and images were also shown at this point. This was followed by a practice run comprising 10 prompted and 10 unprompted images – 50% of which contained targets. The subjects were then given feedback on their performance in the practice run. This should have ensured full familiarisation with the task before the main experiment began.

As part of this training session, the subjects were given an indication of the accuracy level associated with the condition that they were participating in. Rather than a precise summary of the true–positive and false–positive rates, the participants were told whether the number of true–positive prompts was higher, lower or approximately equal to the number of false–positive prompts. The prompting rates associated with the practice run reflected the prompting rates associated with the main study for each condition.

When present, prompts took the form of open circles presented in a blank image to the side of the main stimulus image, as illustrated in figure 8.4. This method of presentation reflected the typical scenario for paper prompting in

mammography, ie: with the prompts presented separate from the original mammograms.

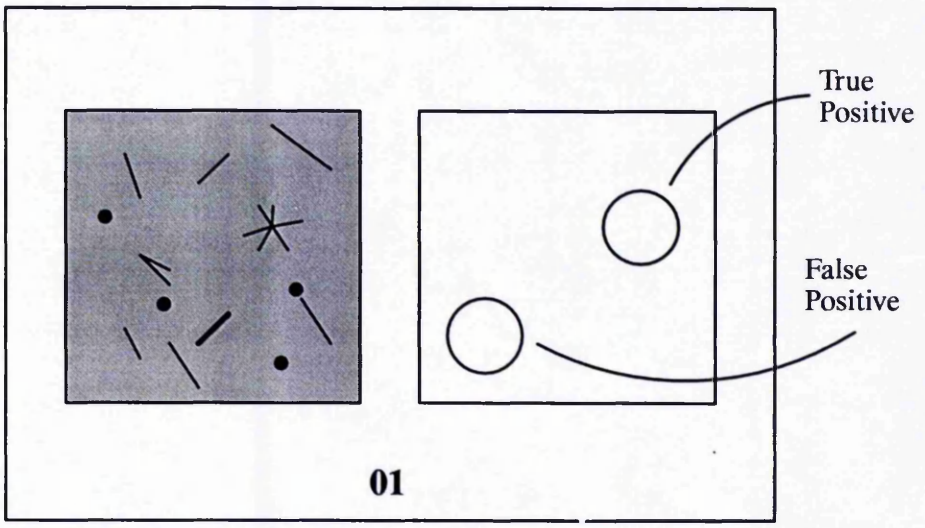

Figure 8.4: Layout of page in prompted cases.

Each of the 9 conditions comprised the same 96 images, and in each condition the same 48 images were assigned to the prompted group. However, the order of presentation of these images was randomised for each subject. Half of the subjects in each condition received the prompted images first and half received the unprompted images first.

8.3 Results and Analysis

Since the experiment took the form of a classic forced-choice signal-detection study, values of d' could be easily calculated for each subject. These values were calculated separately for the prompted and unprompted images. The prompted and unprompted d' scores were then compared for all of the subjects in each

condition. Figure 8.5 shows the overall significance levels (where prompting led to an improvement in detection performance) for each of the experimental conditions.

The first point to note is that prompting was only effective when the number of true–positives was at a certain minimum level (75% true–positive rate in this case). Above this minimum level there appears to be little difference between the higher true–positive rates in terms of the benefits associated with prompting.

		Number of False–positive Prompts		
Number of True–positive Prompts		**8**	**16**	**24**
	16 (100%)	$p<0.01$	$p<0.01$	NS
	12 (75%)	$p<0.05$	$p<0.01$	$p<0.1$
	8 (50%)	NS	*$p<0.05$**	NS

NS = No significant difference

* = Prompting led to reduction in performance

Figure 8.5: Significance levels showing benefits of prompting in each experimental condition

In general it appears that prompting ceases to have a useful effect when the number of invalid prompts is more than 50% higher than the number of valid prompts, though there was a slight improvement when the number of true–positives was 12 and the number of false–positives was 24 (a ratio of 2.0). This improvement, though measurable, was not large enough to be considered significant and is included for the sake of completeness.

One very interesting result is the case where prompting led to a reduction in performance. This condition had a very low (50%) true–positive rate with twice

as many invalid as valid prompts – circumstances that led to a lot of the type of combination errors discussed in chapter 7. It is not really surprising that the prompts reduced performance under these conditions. The time–pressure would have enhanced the role of the invalid prompts as distractors, so that the subject spends much of the limited presentation time checking locations that contain no target, possibly missing the true target in the process, especially if the target is missed by the prompts as it was in 50% of these images.

If this is the case, we might also expect to see the same effect in the next condition where there are three times as many invalid as valid prompts. However, in this case there is no significant difference between the prompted and unprompted images.

One possible explanation for this effect is that the prompts in this case were just too inaccurate and consequently not perceived as being of any benefit to the subject. The subjects may simply have ignored any prompt information so it would not have affected their performance.

This implies that when the prompts are inaccurate, but not hopelessly so, they are still perceived as being useful, even though detection performance is adversely affected. However, when the inaccuracy of the prompts is excessive the perception of benefit is lost and the prompts are ignored.

8.4 Conclusions

This results of this study suggest several interesting points. Firstly, that a minimum level of true–positive accuracy is required in order for prompting to be effective. Secondly, that the number of false–positive prompts should not greatly exceed the number of true–positive prompts or the benefits of prompting will be lost.

Also, it appears that there is a minimum level of prompt accuracy that must not be exceeded in order for prompts to be perceived as useful by the observer. This

suggests that the dangers associated with phenomenon such as combination errors may only occur when the prompting system is accurate enough to be perceived as useful but inaccurate enough to reduce detection performance.

This study has also served to identify the critical range of true–positive /false–positive ratios that should include the point at which prompting effectiveness is lost. This range appears to be between the point at which numbers of true– and false–positives are equal and the point where the number of invalid prompts is double the number of valid prompts. The next experiment, described in chapter 9, will look more closely at this result and determine whether these levels of prompt accuracy are appropriate for a clinical setting.

Chapter 9

Prompting in a Realistic Environment

9.1 Introduction

In order to develop an effective prompting system it is important to understand the ways in which errors in prompt generation may affect the search behaviour of radiologists using the system. The studies described in the last three chapters have shown that excessive numbers of false–positive prompts reduce the effectiveness of prompting. It is therefore important to establish what constitutes an acceptable level of false–positive prompting error so that algorithm developers have a minimum standard of acceptable accuracy for their systems.

The experiment discussed in the previous chapter (chapter 8) described a potential model for the way in which the true–positive and false–positive rates may relate to each other in terms of the information content of the prompts, i.e. the probability that any given prompt was more likely to be valid than invalid. That experiment then went on to identify a critical range in which the effectiveness of a prompting system may break down due to an excessive proportion of invalid prompts.

This previous study, employing a simulated mammography task and non–radiologists, suggested that there is a relationship between the acceptable

false–positive rate and the true–positive rate in an effective prompting system. The number of false–positives should not greatly exceed the number of true–positives, so that any given prompt is as likely to be valid as invalid. Under these conditions the prompts are both perceived as useful and demonstrate an improvement in detection performance.

The highly artificial nature of this task does limit its applicability to prompting in a clinical screening environment. However, the results do make it possible to focus the range of a larger scale study to a size where it may be performed using real mammograms and radiologists.

9.1.1 Objectives

The aim of this final study was to investigate whether the effects of the relationship between true–positive and false–positive rates investigated in the previous artificial study still apply in a realistic environment.

To this end the experiment was carried out in various screening centres within the UK, involving experienced mammographic radiologists viewing films on equipment that would typically be used for mammographic screening.

Three different levels of false–positive prompting rate were used at a fixed true–positive rate. These accuracy levels were chosen to reflect the critical range identified by the simulated mammography experiment described in chapter 7.

The results of this investigation should identify the level of accuracy that is required for a prompting system to be effective in a clinical environment.

9.2 Experimental Method

9.2.1 Images

The data set for the study comprised 100 pairs of mammograms taken from routine screening. Of these, 20 contained subtle malignancies of various types – microcalcifications, spiculated lesions, distortion etc. These malignancies were deliberately selected to be very subtle and difficult to detect. The remaining images consisted of normals and benign structures. The abnormality rate of 20% is somewhat higher than would normally be expected during a screening session, which might contain around 5–10% abnormalities. However, a rate of 20% does allow for a reasonably high number of abnormalities that can be investigated, while still simulating the typical screening situation in which the great majority of cases are normal.

The true–positive prompting rate was fixed at around 90%, since this is a reasonable goal to expect from a prompting system. There were 3 false–positive rates, each of which corresponded to a different experimental condition. These rates were determined by the ratio of the number of false–positive prompts (nFP) to the number of true–positive prompts (nTP). The three experimental conditions were selected to have nFP/nTP = 1.0, 1.5 and 2.0. Based on the results of the 'simulated mammogram' study (chapter 8), it was expected that the first condition should lead to effective prompting and the third condition should not. Whether the second condition (nFP/nTP = 1.5) would lead to improved performance was not clear.

The data for the experiment included abnormalities for which no effective computer–based detection system has yet been developed. There were also highly specific requirements for the sensitivity and specificity of the prompts. For these reasons, the prompts used in the experiment were simulated and randomly assigned to the appropriate images.

9.2.2 Subjects

In total, 30 experienced radiologists from 11 screening centres in the UK took part in the study. All of these participants were either consultants or senior registrars and all were involved in mammographic screening on a regular basis. Each subject was assigned to one of 3 experimental conditions. This assignment was random, but for practical reasons all participants in a given centre were assigned to the same condition.

9.2.3 Procedure

The procedure for this study was very similar to that described in chapter 7. Each radiologist was presented with the original film pairs presented on standard viewing equipment in their screening centre. Each film pair was accompanied by a hardcopy of the digital mammograms. This hardcopy contained prompt information when it was available and also served as a response form on which the radiologists were able to record their judgements. Appendix 1 contains examples of both prompted and unprompted hardcopies.

Each radiologist was presented with 100 film pairs, 50 of which were prompted and 50 unprompted. The malignancies were equally distributed between the prompted and unprompted images. Half of the subjects in each condition saw the prompted films first and half saw the unprompted first. Film pairs were randomly assigned to either the prompted or unprompted groups and for each screening centre this assignment was different.

Apart from the copy mammograms, each hardcopy contained a six–point rating scale ranging from '0: Normal' to '5: Malignant' as described in section 7.2.3. Also present was a list of recommended further actions chosen to reflect the typical courses of action that a radiologist would recommend in practice. It should be noted that due to differences in working practices at different screening centres

it is not practical to develop a definitive list of possible recommendations, though the options listed here are among the more common alternatives:

Craniocaudal view

Compression

Magnification

Ultrasound

Surgical Opinion

No Further Action

Each participating radiologist was asked to locate any abnormalities in the original mammograms, mark the locations on the hardcopy and provide a rating of the severity of case using the rating scale on the hardcopy. They were also asked to mark which course(s) of action on the list provided they would recommend in each case.

When prompts were presented they should have been used in conjunction with the original films to locate potentially suspicious regions. Prior to the experiment, the radiologists were told the approximate accuracy levels of the prompting system, so that they knew what to expect rather than alter their strategy after a few films because they have made their own decisions about accuracy levels. They were also informed of the ratio of normal to abnormal films.

The radiologists were not given any practice runs before the main session as the task should have been familiar enough to make it unnecessary. However, in order to explain what was expected of them, they were shown a few examples of hard copies both with and without prompts when the experimental task was explained to them. None of the examples included films that appeared in the experiment.

9.3 Results and analysis

Truth data was obtained from an experienced consultant with access to patient records and pathology data in addition to the original films. All of the cases marked as malignancies had pathological confirmation of the diagnosis. Normal and benign cases were taken from screening sessions conducted some years prior to this experiment to ensure that no interval cancers had arisen in the interim.

Each of the three conditions was analysed separately. The detection performance of each radiologist under both the prompted and control conditions was determined by means of ROC analysis. This involved calculating the number of true–positive and false–positive responses at each point of the rating scale, since these points represent different levels of response bias. These values were then plotted to yield an ROC curve for each of the subjects. The individual ROC curves were then pooled to yield the overall curves by averaging the numbers of true–positives and false–positives at each criterion level.

9.3.1 Condition 1

In this condition the numbers of true–positive and false–positive prompts were equal. Figure 9.1 shows the pooled ROC curves for the 10 subjects assigned to this condition, while figure 9.2 lists the values of A_z for each of the radiologists under each of the prompted and unprompted conditions (see chapter 3 for discussion of A_z and related measures.)

In figure 9.1 the broken line represents the raw experimental data, while the solid line represents the theoretical ROC curve derived by the software package ROCFIT.

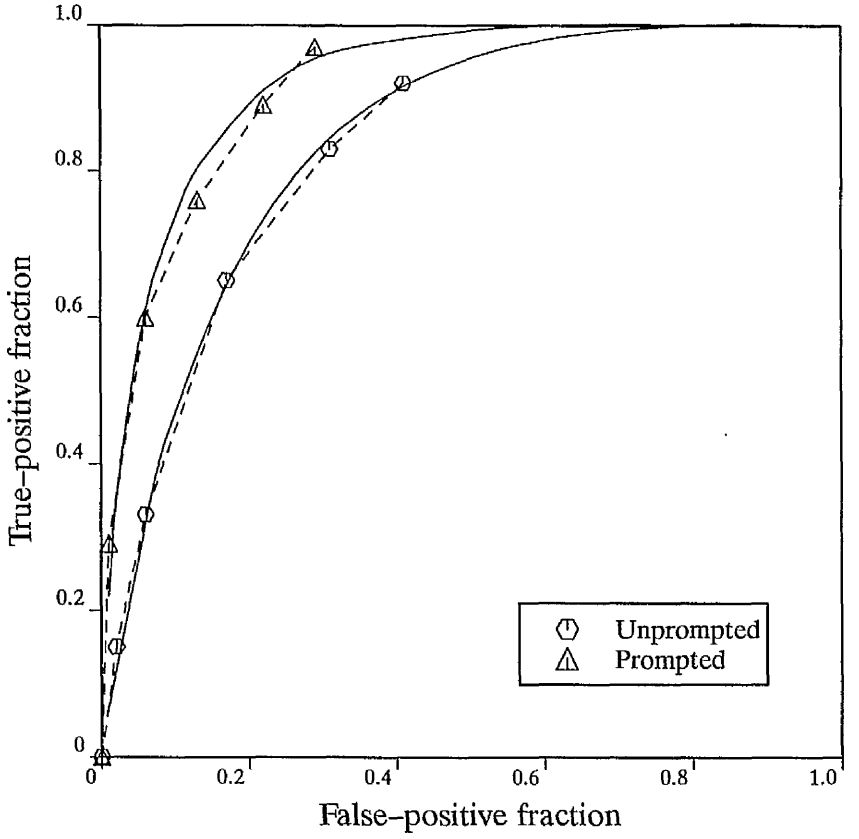

Figure 9.1: ROC curves showing prompted and unprompted performance for prompt level 1.

Subject	Unprompted	Prompted
1	0.8527	0.9067
2	0.8890	0.9357
3	0.7633	0.9197
4	0.8319	0.8794
5	0.8125	0.9163
6	0.8727	0.9326
7	0.9042	0.9064
8	0.9030	0.9541
9	0.7848	0.9503
10	0.7992	0.9215

Figure 9.2: A_z scores of each subject for prompting condition 1.

A related t–test applied to the A_z scores showed in figure 9.2 revealed that prompting led to a significant improvement in performance ($t_{obs} = 4.79$, $p < 0.005$). This suggests that prompting was beneficial to the radiologists at this level of accuracy, as would be predicted by the results of the experiment discussed in chapter 8.

9.3.2 Condition 2.

In this second experimental condition the number of false–positive prompts was 50% greater than the number of true–positive prompts, so that nFP/nTP was 3:2.

Once again the results for the 10 subjects were pooled to generate the pooled ROC curve shown in figure 9.3.

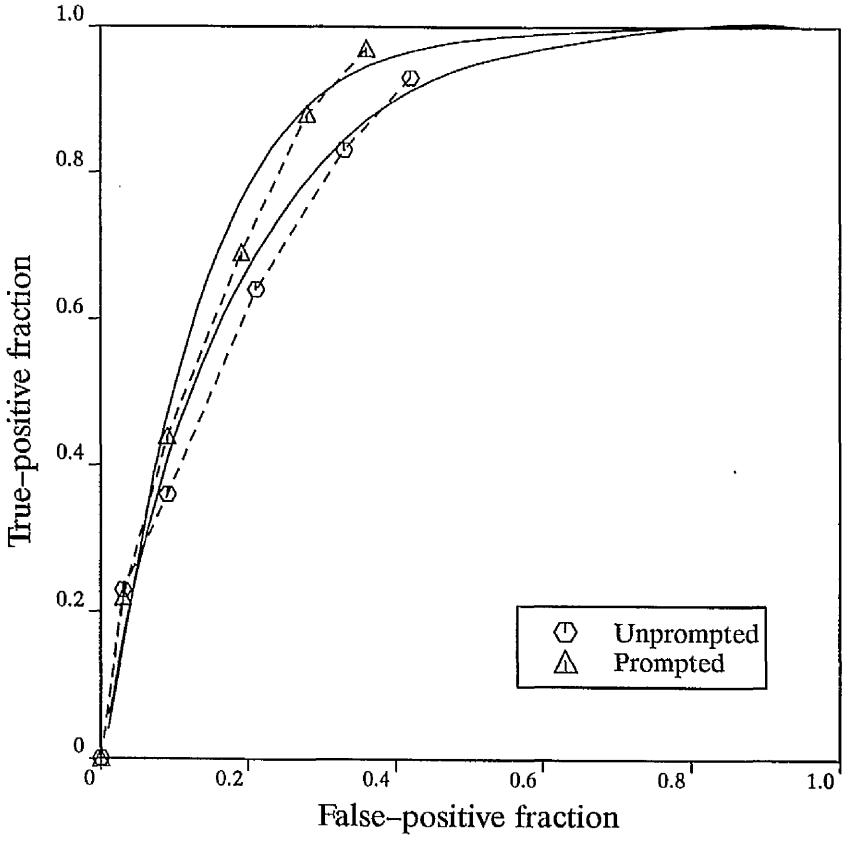

Figure 9.3: Pooled ROC curves for prompted and unprompted cases in prompt condition 2

A_z scores for the individual subjects are shown in figure 9.4. A related t–test applied to these scores showed that prompting led to a significant improvement in performance ($t_{obs} = 2.39$, $p < 0.025$). While the improvement in this case is not as marked as that observed in the level 1 condition, the results suggest that prompting was still beneficial to the radiologists at this level of accuracy.

Subject	Unprompted	Prompted
1	0.8446	0.8963
2	0.7854	0.8546
3	0.8570	0.9012
4	0.8088	0.8833
5	0.7900	0.8841
6	0.7717	0.8551
7	0.8619	0.8225
8	0.8604	0.8303
9	0.8390	0.8448
10	0.8217	0.8472

Table 9.4: A_z scores of each subject for prompting level 2.

9.3.3 Condition 3

In this condition the number of false–positive prompts was double the number of true–positive prompts. Figure 9.5 shows the pooled ROC curves for the 10 subjects assigned to this condition, while figure 9.6 lists the values of A_z for each of the radiologists under each of the prompted and unprompted conditions.

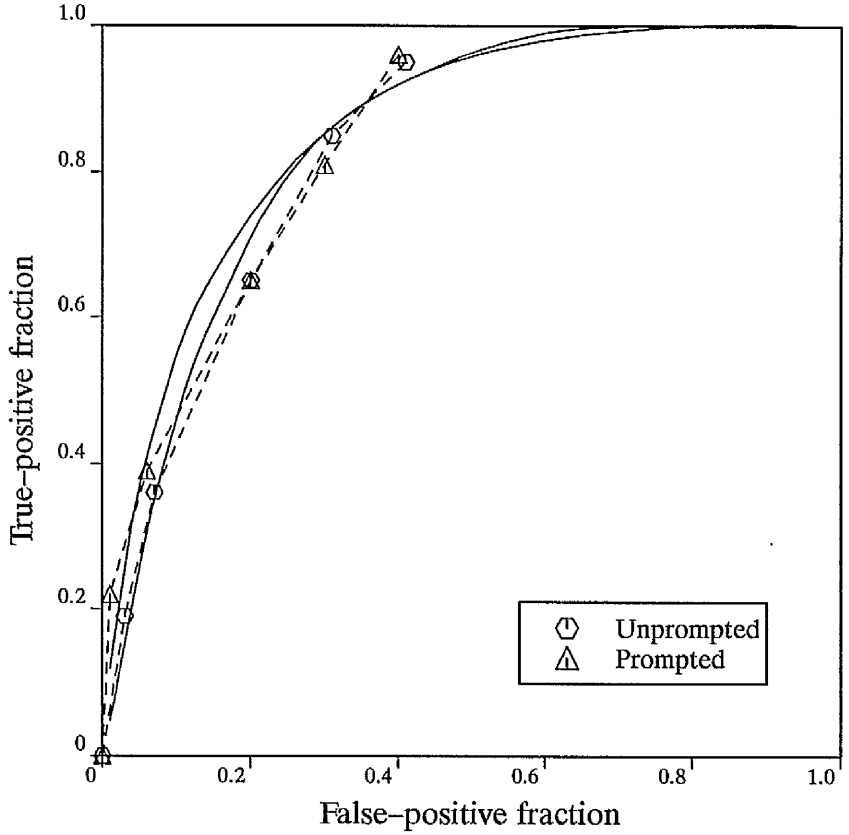

Figure 9.5: Pooled ROC curves for prompted and unprompted cases in prompt level 3

Subject	Unprompted	Prompted
1	0.8024	0.7818
2	0.7296	0.8816
3	0.9350	0.7722
4	0.9418	0.9214
5	0.8567	0.8940
6	0.8203	0.7989
7	0.8776	0.8608
8	0.8796	0.8373
9	0.8546	0.8740
10	0.8116	0.9050

Table 9.6: A_z scores of each subject for prompting level 3.

An analysis of the A_z scores for this final condition shows that there is no significant difference between the prompted and unprompted cases in prompt condition 3 ($t_{obs} = 0.07$). This suggests that prompting ceases to be beneficial to the radiologist at this level of accuracy.

9.3.4 Comparison of difficulty levels.

Each clinic in the study was presented with a different randomised order of films. This meant that a film that was prompted for one radiologist might be unprompted for another. In order to establish whether different film presentation orders had increased or decreased the difficulty of the task, the unprompted performance for the 3 conditions was compared.

There were found to be no significant differences in performance between the unprompted cases in level 1 and level 3 ($t_{obs} = 0.54$), level 2 and level 3 ($t_{obs} = 0.68$) or level 1 and level 2 ($t_{obs} = 1.02$). This supports the suggestion that the observed performance improvement in the level 1 and level 2 conditions was due to prompting and not due to differences in the degree of difficulty between the prompted and unprompted cases.

9.3.5 Analysis of 'further action' results

The radiologists' responses to the 'recommended further action' task were also analysed separately for each experimental condition. The number of each type of response was counted for each radiologist in each of the prompted and unprompted conditions. The results were further subdivided into normal and abnormal cases – the latter being those cases known to contain a malignancy.

Figure 9.7 shows the numbers of each type of response given by the radiologists in experimental condition 1.

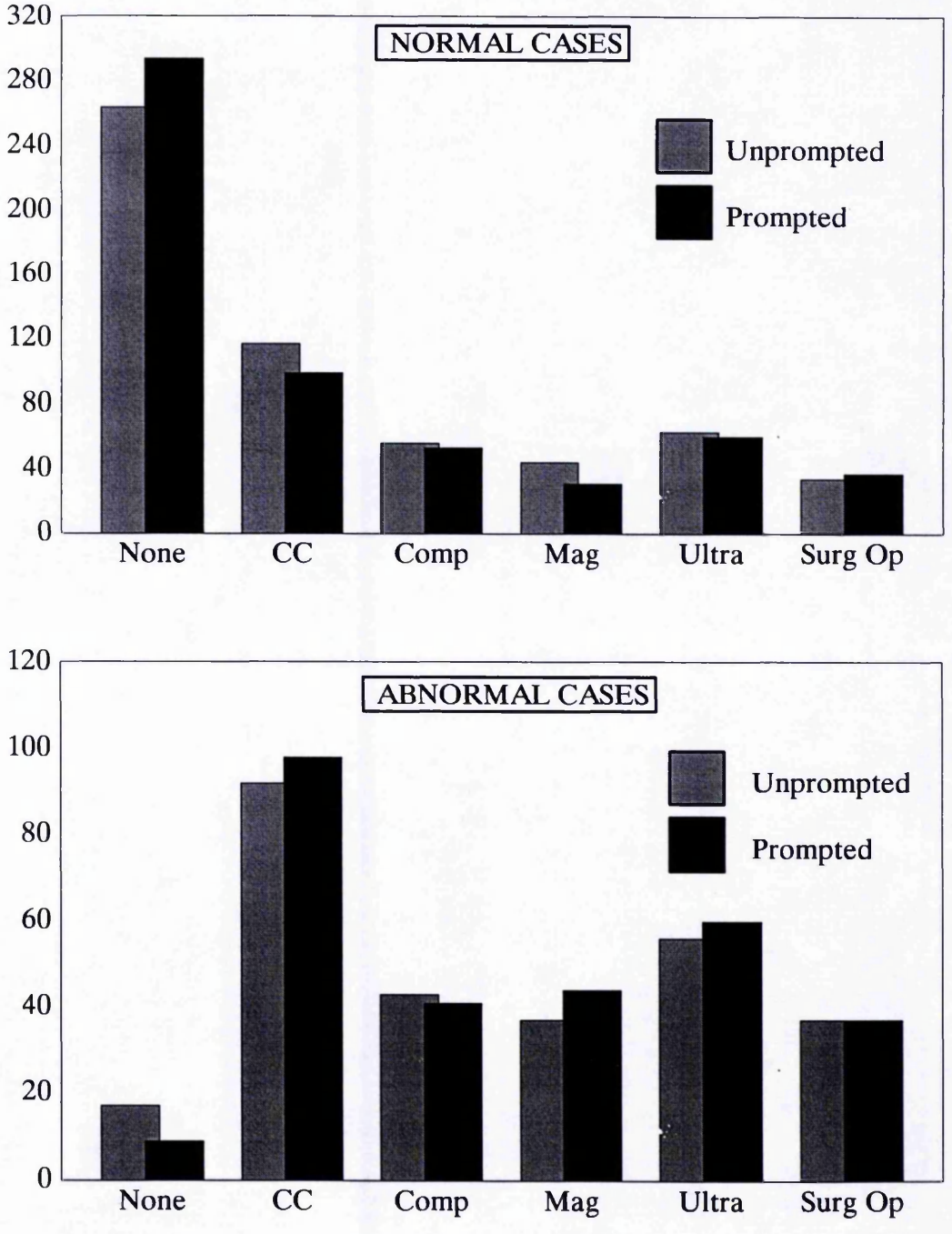

Figure 9.7: Recommendations for further action in condition 1.

In the prompted condition significantly fewer normal cases were recommended for further action ($t_{obs} = 3.40, p < 0.005$). There was no significant difference in the

number of abnormal cases recommended for no further action ($t_{obs} = 1.64$). This suggests that prompting has reduced the number of false recalls without significantly affecting the number of genuine recalls.

Figure 9.8 shows the further action responses given by all of the radiologists in condition 2.

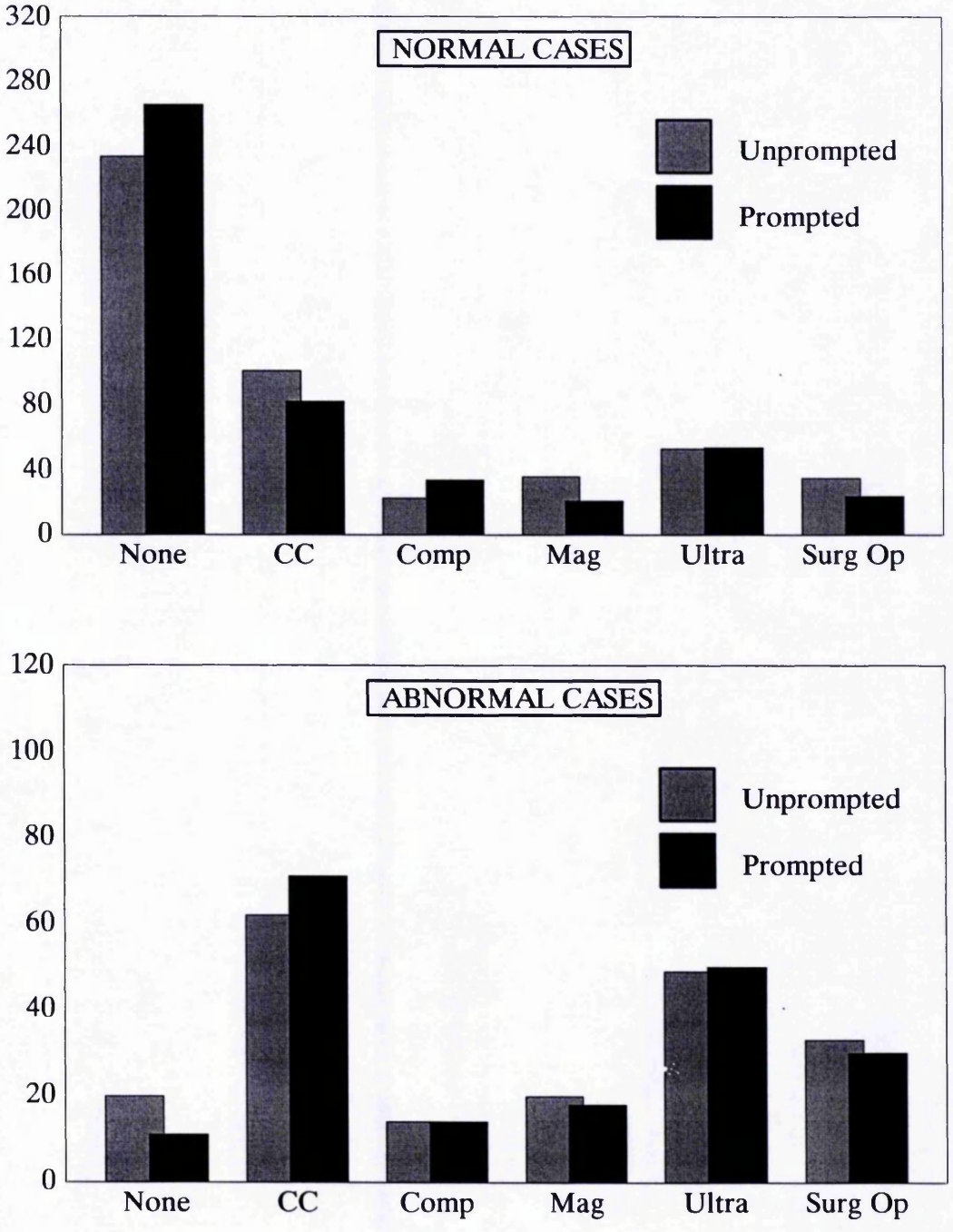

Figure 9.8: Recommendations for further action in condition 2.

As with prompt level 1, in the prompted condition significantly fewer normal cases were recommended for further action ($t_{obs} = 3.38$, $p < 0.005$). Again there was no significant difference in the number of abnormal cases recommended for no further action ($t_{obs} = 2.00$). As with prompt level 1, prompting has reduced the number of false recalls without significantly affecting the number of genuine recalls.

Figure 9.9 shows the responses for recommended further action in the final condition, where the number of false–positives was double the number of true–positives.

For this level of accuracy prompting does not lead to any significant change in the number of cases recommended for no further action, either for normal ($t_{obs} = 0.99$) or for abnormal ($t_{obs} = 0.69$) cases. This suggests that prompting had no effect on the decision to recall at this level of accuracy.

It should be noted that any analysis of the different types of further action recommended beyond the figures for 'no further action' will be unreliable, as the different clinics involved in the study all have different working practices. Many participating radiologists commented that the options in the study did not reflect the range of options they would normally have available.

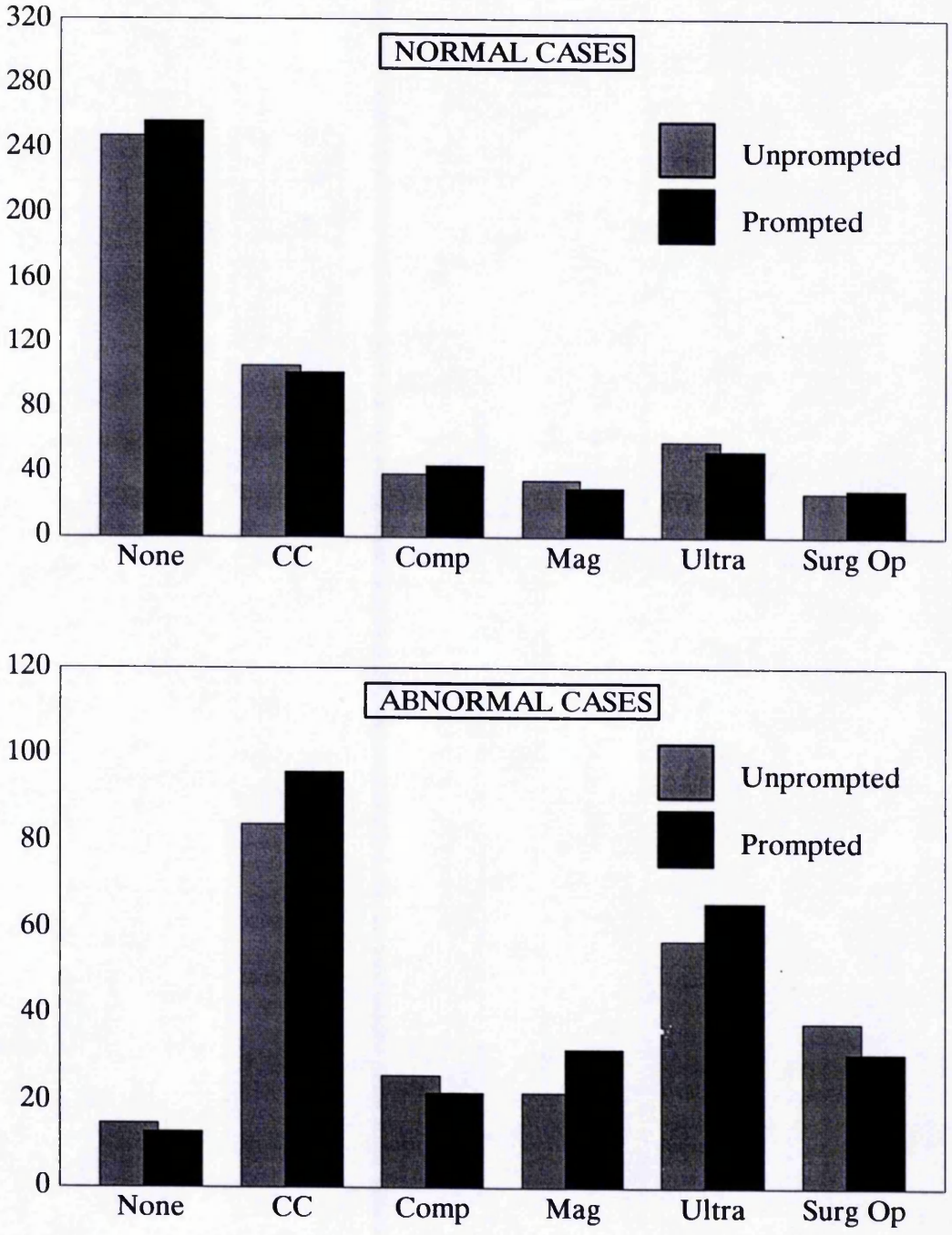

Figure 9.9: Recommendations for further action in condition 3.

9.4 Summary and Conclusions

The experiment described in the preceding sections is the largest scale study of prompting effectiveness in a realistic environment to date.

Once again, prompting has been demonstrated to be an effective aid to the radiologist providing that the accuracy level of the prompts is sufficient.

In this experiment, both condition 1 and condition 2 showed an improvement in the detection performance of the radiologists. The improvement was much greater in the first condition, where the numbers of true–positive and false–positive prompts were equal, than in the second condition where there were 50% more invalid than valid prompts. In the third condition, where the number of false–positive prompts was double the number of true–positive prompts, there appeared to be no benefit from prompting.

It is interesting to note that in the two conditions in which prompting was effective, there was also a significant reduction in the number of normal cases recommended for further action (cases that would be recalled in practice) in the prompted condition. However, there was no significant change in the number of abnormal cases recommended for further action. This suggests that prompting is improving performance by reducing the number of invalid judgements made by the radiologist.

The results of this study suggest that prompting is still effective as long as the numbers of false–positive prompts do not exceed the numbers of true–positive prompts by more than 50%. However, it should be noted that the proportion of abnormalities in the data–set was higher than would normally be seen during screening. Given the lower proportion of abnormalities that would be present in a clinical environment and the greater number of images that would be studied over time, it is possible that the acceptable false–positive rate might need to be lower than this. For this reason a more acceptable limit on prompting accuracy

might be that the number of false–positive prompts does not exceed the number of true–positive prompts generated by the detection system.

Chapter 10

Summary and Conclusions

This thesis began with a discussion of the problem of breast cancer and the notion of early detection as one of the most effective means available for its control. This early detection and treatment requires an imaging technique that can detect the subtle signs of early breast cancer and consequently mammography has been adopted as the standard technique for screening. The effectiveness of mammography critically depends on the ability of radiologists to detect these subtle abnormalities embedded in the complex background associated with mammograms.

One possible application of computer-based imaging techniques to mammography is the development of a prompting system that automatically detects abnormalities in digital mammograms and points out their locations to the radiologist, thus aiding the detection of these abnormalities. A number of approaches to the automatic detection of common mammographic abnormalities have been discussed. Two such techniques have been described in detail.

The remainder of this thesis described several psychophysical studies designed to investigate the effectiveness of prompting and to study the effects of errors in

prompt generation on the detection performance of radiologists working with a prompting system.

The experiments described in this thesis have demonstrated that under the right conditions, prompting can be a useful aid to the radiologist in the detection of subtle mammographic abnormalities. In a large scale realistic study (chapter 9) employing 30 experienced radiologists and 100 film pairs containing a variety of abnormalities, prompting led to an increase in detection performance, provided the accuracy of prompt generation was high enough.

An important observation from the experiments in this study is that false–positive prompts play a critical role in the effectiveness of a prompting system. The investigations described in chapters 6 and 9 both showed that prompting ceases to be a useful aid to the radiologist once the false–positive rate becomes too high. Chapters 8 and 9 have also suggested that an acceptable level of false–positive prompt generation can be set by not allowing the number of false–positive prompts to greatly exceed the number of true–positive prompts.

Taken individually, the results of the studies described in chapters 8 and 9 both support the notion of an important relationship between the acceptable numbers of true–positive and false–positive prompts. The fact that these two studies were very different in nature and were carried out independently of one another lends even greater support to this conclusion.

It is not a trivial matter to ensure that a prompting system generates less false–positives than true–positives. Typically, about 5% of screening mammograms contain some form of abnormality, and perhaps 90% of these are detected by a prompt generation system. This means that the system should generate a false–positive prompt no more often than once in every twenty cases. At present no reported algorithm comes close to this level of accuracy for the

detection of a single class of abnormality. Certainly nothing is available that can attain these accuracy levels on a range of abnormalities.

It appears that although prompting may be an effective technique in theory, its utility as a practical aid to radiologists is restricted by the performance of the prompt generation algorithms. Unfortunately these algorithms have not yet reached a stage where they are accurate or fast enough for real-time prompting of screening mammograms to be a realistic proposition.

The technology for capturing primary digital mammograms, with no need for the use of conventional film, is becoming more and more common. As such equipment increasingly becomes the standard for screening mammography, the feasibility of prompting and other computer-based analysis techniques becomes more apparent. The process of digitising mammograms so that they are accessible to computer vision techniques is costly and fraught with problems (see section 4.1.1). Removal of this step should make the application of computer vision techniques to mammography much more realistic.

References

Anderson DE (1974) "Genetic Study of Breast Cancer: Identification of a High Risk Group", Cancer: 34, 1090.

Asbury DL (1990) "The UK Breast Cancer Screening Programme", in "Manchester Breast Screening Manual", University Dept of Medical Illustration, Withington Hospital, Manchester.

Astley SM & Taylor CJ (1990) "Combining Cues for Mammographic Abnormalities", Proc BMVC 1990, 253–258.

Austoker J & Humphries J (1988) "Breast Cancer Screening", CRC Practical Guides for General Practice: 6, Oxford University Press.

Bailar JC (1978) "Mammographic Screening: A Reappraisal of Benefits and Risks.", Clin. Obstet. Gynec.: 21, 1–14.

Baral E, Larrson LE, Mattson B (1977) "Breast cancer following irradiation of the breast", Cancer:4, 2905–10.

Baum (1988) "Breast Cancer: the facts", New York, OUP.

Berg (1984) "Clinical Implications of Risk Factors of Breast Cancer", Cancer: 53, 589–591.

Bourrely C & Muller S (1990) "Detection of Microcalcifications in Mammographic Images", in Fogelman Soulie F & Herault J (eds) "Neurocomputing", NATO ASI Series, Vol F68, Springer–Verlag, Berlin, 325–328.

Breslow L, Thomas LB & Upton AC (1977) "Final reports of ad hoc working groups on mammographic screening for breast cancer and a summary of their joint findings and recommendations", J. Nat Cancer Inst:59, 467.

Brzakovic D, Luo XM & Brzakovic P (1990) "An approach to automated detection of tumours in mammograms", IEEE Trans Med Imaging: 9(3), 233–41.

Burns PE (1978) "False-negative Mammograms and Delay in the Diagnosis of Breast Cancer", New England Med J: 299, 201–2.

Burt PJ (1981) "Fast filter transforms for image processing" Computer Vision, Graphics and Image Processing: 16, 20–51.

Chan H–P, Doi K, Vyborny CJ, Schmidt RA, Metz CE, Lam KL, Ogura T, Wu Y & Macmahon H (1990) "Improvement in Radiologists' Detection of Clustered Microcalcifications on Mammograms: The Potential of Computer–Aided Diagnosis", Invest Radiol: 25, 1102–1110.

Chan H–P, Doi K, Vyborny CJ, Lam K–L & Schmidt RA (1988a) "Computer–Aided Detection of Microcalcifications in Mammography: Methodology and Preliminary Study", Invest Radiol: 23, 664–671.

Chan H–P, Doi K, Lam K–L, Vyborny CJ, Schmidt RA & Metz CE (1988b) "Digital Characterisation of Clinical Mammographic Microcalcifications: Applications in Computer–Aided Detection", SPIE Vol 914 Medical Imaging II, 591–592.

Chan H–P, Vyborny CJ, Macmahon H, Metz CE, Doi K & Sickles EA (1987a) "Digital Mammography: ROC Studies of the Effects of Pixel Size and Unsharp–mask Filtering on the Detection of Subtle Microcalcifications", Invest Radiol: 22, 581–589.

Chan H–P, Doi K, Galhotra S, Vyborny CJ, Macmahon H & Jokich PM (1987b) "Image Feature Analysis and Computer–Aided Diagnosis in Digital Radiography: 1. Automated Detection of Microcalcifications in Mammography", Med Phys: 14(4), 538–548.

CRC Factsheet (1988) "Breast Cancer", Facts on cancer, Cancer Research Campaign, London.

Davies DH & Dance DR (1990) "Automatic Computer Detection of Clustered Calcifications in Digital Mammograms", Phys Med Biol: 35(8), 1111–1118.

Dhawan AP, Buelloni G & Gordon R (1986) "Enhancement of Mammographic Features by Optimal Adaptive Neighbourhood Image Processing", IEEE Trans Med Imaging: MI–5(1), 8–15.

Downing CJ & Pinker S (1985) "The Spatial Structure of Visual Attention" in Posner MI & Marin OSM (eds) "Attention and Performance XI", Hillsdale NJ, LEA.

Ellman R, Angeli N, Chritians A Moss Chamberlain J & Maguire P (1989) "Psychiatric Morbidity Associated with Screening for Breast Cancer", Br J Cancer 60: 143–149.

Eriksen BA & Eriksen CW (1974) "Effects of noise letters on the identification of a target letter in a nonsearch task", Perception & Psychophysics: 16, 143–149.

Eriksen CW & Murphy TD (1987) "Movement of Attentional focus across the visual field: A critical look at the evidence", Perception & Psychophysics: 42, 299–305.

Eriksen CW & Yeh Y–Y (1985) "Allocation of Attention in the Visual Field" J Exp Psych: Human Perception & Performance: 11, 583–597.

Fam BW & Olson SL (1988) "The Detection of Calcification Clusters in Film–Screen Mammograms: A Detailed Algorithmic Approach". SPIE Vol 914, Medical Imaging II, 620–34.

Fam BW, Olson SL, Winter PF & Scholz FJ (1988) "Algorithm for the Detection of Fine Clustered Calcifications on Film Mammograms", Radiology: 169, 333–337.

Fechner GT (1860) "Elemente der Psychophysik", Leipzig, Breitkopf & Hartel. Reissued 1964 by Bonset, Amsterdam.

Feig SA (1986) "Screening Mammography: Benefits and Risks" in Moskowitz M (ed) "Diagnostic Categorical Course in Breast Imaging", 75–84.

Feig SA et al (1977) "Thermography, mammography and Clinical Examination in Breast Cancer Screening", Radiology: 122, 123.

Fisher ER (1985) "What is Early Breast Cancer ?", in Zander J & Baltzer J (eds) "Early Breast Cancer: Histopathology, Diagnosis & Treatment", Berlin, Springer–Verlag, 1–13.

Fornage BD, Toubas O & Morel M (1987) "Clinical, Mammographic and Sonographic Determination of Preoperative Breast Cancer Size",Cancer: 60, 765–71.

Forrest (1986) "Breast Cancer Screening, HMSO.

Gale AG, DeSilva ES, Walker GE, Roebuck EJ & Worthington BS (1989) "Vigilance Decrement and Radiological Reporting" in Megaw E (ed) "Contemporary Ergonomics", Taylor & Francis.

Gale AG, Johnson F & Worthington BS (1979) "Psychology and Radiology", in Oborne DJ, Gruneberg M & Eiser "Research in Psychology & Medicine", Academic Press.

Gale AG & Worthington BS (1986) "Visual Attention in Diagnostic Radiology", in Oborne DJ, "Contemporary Ergonomics", Burlington, Taylor & Francis.

Gallagher HS (1985) "Pathogenesis of Early Breast Cancer", in Zander J & Baltzer J (eds) "Early Breast Cancer: Histopathology, Diagnosis & Treatment", Berlin, Springer–Verlag, 14–19.

Gallagher HS & Martin JE (1971) "An Introduction to the Concept of Minimal Breast Cancer", Cancer: 28, 1505–1507.

Giger ML, Yin FF & Doi K (1990a) "Image features in mammographic masses used in the development of computerised schemes", in Arenson R & Friedenberg RM (eds) "SCAR 90: Computer applications to assist radiology".

Giger ML, Yin FF, Doi K, Metz CE, Schmidt RA & Vyborny CJ (1990b) "Investigation of methods for the computerised detection and analysis of mammographic masses", SPIE Vol 1233: Medical Imaging IV – Image Processing. 183–185.

Green DM & Swets JA (1966) "Signal Detection Theory and Psychophysics", New York, Wiley. Reprinted 1974 by Krieger, Huntington, NY.

Gregg EC (1977) "Radiation Risks with Diagnostic X–rays", Radiology: 123, 447.

Haffty BG, Kornguth P, Fischer D, Beinfield M & McKhann C (1991) "Mammographically Detected Breast Cancer: Results with Conservative Surgery and Radiation Therapy", Cancer: 67(11), 2801–2804.

Helman P (1977) "Whither Breast Cancer ?: Report of the Inaugural Meeting of the National Breast Cancer Group" S Afr Med J: 52, 711–713.

Helmholtz H (1866/1925) "Physiological Optics", Sothall JPC (ed), New York, Dover.

Henderson BE, Pike MC & Ross RK (1984) "Epidemiology and Risk Factors" in Bonadonna G (ed) "Breast Cancer: Diagnosis and Management", John Wiley & Sons,15-33.

Hopwood P & Maguire P (1990) "The Psychological Impact of the Diagnosis and Treatment of Breast Cancer", in "Manchester Breast Screening Manual", University Dept of Medical Illustration, Withington Hospital, Manchester.

Howell DC (1987) "Statistical Methods for Psychology", Duxbury Press, Boston.

Hoyer A & Spiesberger W (1979) "Computerised Mammogram Processing", Phillips Tech Rev: 38.

Humphries GW (1984) "Shape Constancy: The Effects of Changing Shape Orientation and the Effects of Changing Focal Features", Perception & Psychophysics: 36, 50-64.

Humphries GW & Bruce V (1989) "Viual Cognition: Computational, Experimental and Neuropsychological Perspectives" Hillsdale NJ, LEA.

Hutt IW (1992) "The Effects of Prompting on the Detection of Clustered Microcalcifications in Digital Mammograms", MSc Thesis, University of Manchester.

James W (1890/1950) "The Principles of Psychology 1" New York, Dover.

Karssemeijer N, Frieling JTM & Hendricks JHCL (1993) "Spatial Resolution in Digital Mammography", Invest Radiol: 28, 413-9.

Kegelmeyer WP (1992) "Computer Detection of Stellate Lesions in Mammograms", SPIE Vol 1660, Biomedical Image Processing & 3D Microscopy, 446-54.

Kopans DB (1987) "Nonmammographic breast imaging techniques: current status and future developments", Radiol Clin North Am: 25, 961-971.

Kundel HL & Nodine CF (1978) "Studies of Eye Movements and Visual Search in Radiology" in Seders JAW, Fisher D & Monty R (eds) "Eye Movements and the Higher Psychological Functions", Hillsdale NJ, LEA.

Kundel HL, Nodine CF &Carmody DP (1978) "Visual Scanning, Pattern Recognition & Decision Making in Pulmonary Nodule Detection", Invest Radiol: 13, 175–181.

Lai S–M, Li X & Bischof WF (1989) "On techniques for detecting circumscribed masses in mammograms", IEEE Trans Med Imaging: 8(4), 377–86.

Lai S–M, Li X & Bischof WF (1988) "Automated Detection of Breast Tumours", in Krzyzak et al (eds) "Computer Vision and Shape Recognition", 115–32.

Lanyi M (1985) "Morphological Analysis of Microcalcifications", in Zander J & Baltzer J (eds) "Early Breast Cancer: Histopathology, Diagnosis & Treatment", Berlin, Springer–Verlag.

Lau T–K & Bischof WF (1991) "Automated Detection of Breast Tumours using the Asymmetry Approach", Computers and Biomedical Research: 24, 273–295.

Lesnick GJ (1977) "Detection of Breast Cancer in Young Women", J AM Med Assoc: 237, 967–9.

Letton AH, Wilson JP & Mason EM (1977) "The value of breast screening in women less than 50 years of age", Cancer: 40, 1–3.

Lissner J, Kessler M, Anhalt G, Hahn D Wendt T & Seiderer M (1985) "Developments in Methods for Early Detection of Breast Cancer" in Zander J & Baltzer J (eds) "Early Breast Cancer: Histopathology, Diagnosis & Treatment", Berlin, Springer–Verlag. 93–112.

Lloyd P, Mayes A, Manstead ASR, Meudell PR & Wagner HL (1984) "Introduction to Psychology: An Integrated Approach", London, Fontana.

Luce RD (1963) "A threshold theory for simple detection experiments" Psychological Review: 70, 61–79.

MacMahon B, Cole P, Brown J (1973) "Etiology of Human breast Cancer: A Review", J Nat Cancer Inst: 50, 21–42.

MacMillan NA & Creelman CD (1991) "Detection Theory: A User's Guide", Cambridge, Cambridge University Press.

McClelland R (1990) "Earlier detection of Breast Cancer: an overview", in Brunner S, Langfeldt B (eds) "Recent Results in Cancer Research vol 119: Advances in Breast Cancer Detection", Springer-Verlag, Berlin, 10-17.

Miller P & Astley S (1993) "Automated Detection of Breast Asymmetry using Anatomical Features", IJPRAI special issue on mammographic image analysis.

Morgan RH (1981) "Benefit-Risk Ratios in Mammography" in Lewinson EF & Montague ACW (eds) "Diagnosis and Treatment of Breast Cancer", Baltimore, Williams & Wilkins.

Moskowitz M (1977) "Screening for Breast Cancer", J AM Med Assoc: 238, 213.

Muir BB, Kirkpatrick AE, Roberts MM, Duffy SW (1984) "Oblique-view mammography: adequacy for screening. Work in Progress. Radiology: 151, 39-41.

Muller JWT, van Waes PFGM, Koehler PR (1983) "Computed tomography of breast lesions: comparison with X-ray mammography", J Comp Assist Tomogr: 7, 650-654.

Murphy WA & DeSchryver-Kecskemeti K (1978) "Isolated Clustered Microcalcifications in the Breast: Radiologic-Patholgic Correlation", Radiology:127, 335-341.

Nab HW, Karssemeijer N, Van Erning LJTHO & Hendricks JHCL (1992) "Comparison of Digital and Conventional mammography: ROC study of 270 mammograms", Med Inform: 17(2), 125-131.

Nishikawa RM, Giger ML, Doi K, Vyborny CJ & Schmidt RA (1993) "Computer-aided detection and diagnosis of masses and clustered microcalcifications from digital mammograms", SPIE symposium on Electronic Imaging.

Olson SL, Fam BW, Winter PF, Scholz FJ, Lee AK & Gordon SE (1988) "Breast Calcifications: Analysis of Imaging Properties", Radiology: 169, 329-332.

Posner MI (1978) "Chronometric Explorations of Mind" Hillsdale NJ, LEA.

Posner MI (1980) "Orienting of Attention" Quart J Exp Psych: 32, 3-25.

Rich E & Knight K (1991) "Artificial Intelligence", McGraw-Hill Inc.

Robson C (1983) "Experiement, Design & Statistics in Psychology", Penguin Books, Middlesex.

Roebuck EJ (1986) "Mammography and screening for breast cancer" Br Med J: 292, 223-226.

Roebuck EJ (1990) "Clinical Radiology of the Breast", Heinemann Medical Books, Oxford.

Schulman AI & Mitchell RR (1966) "Operating Characteristics from yes-no and Forced-choice Procedures", Journal of Acoustical Science of America: 40, 473-477.

Semmlow JL, Shadagopappan A, Ackerman LV, Hand W & Alcorn FS (1980) "A fully automated system for screening xeromammograms", Computers and Biomedical Research: 13, 350-62.

Shapiro S, Strax P, Venet L& Venet W (1973) "Changes in a 5 year breast cancer mortality in a breast screening program", in Lippincott JB (ed) "Proc of 7th Nat Cancer Conf", Philadelphia.

Shulman GL, Remington RW & Mclean JP (1979) "Moving Attention through Visual Space", J Exp Psych: Human Perception & Performance: 5, 522-526.

Sickles EA (1990a) "One versus Two Views per Breast for Screening" in Brunner S, Langfeldt B (eds) "Recent Results in Cancer Research vol 119: Advances in Breast Cancer Detection", Springer-Verlag, Berlin, 81-87.

Sickles EA (1990b) "Imaging techniques other than mammography for the detection and diagnosis of breast cancer", in Brunner S, Langfeldt B (eds) "Recent Results in Cancer Research vol 119: Advances in Breast Cancer Detection", Springer-Verlag, Berlin, 127-135.

Sickles EA (1982) "Mammographic Detectability of Breast Calcifications", Am J Roent: 139, 913–918.

Sickles EA, Weber WN, Galvin HB, Ominsky SH & Sollitto RA (1986a) "Baseline screening mammography: one vs two views per breast", AJR: 147, 1149–1153.

Sickles EA, Weber WN, Galvin HB, Ominsky SH & Sollitto RA (1986b) "Mammographic Screening: How to Operate Successfully at Low Cost", Radiology: 160, 95–97.

Simpson AJ & Fitter MJ (1973) "What is the best index of detectability ?", Psychological Bulletin: 80, 481–488.

Spiesberger W (1979) "Mammogram Inspection by Computer", IEEE Trans Biomedical Engineering: 26(4).

Spiesberger W & Groh G (1977) "Outlining of Microcalcifications by Computer-Assisted Mammogram Inspection", Medicamundi, 22(3).

Strax P (1981) "Mammography"in Lewinson EF & Montague ACW (eds) "Diagnosis and Treatment of Breast Cancer", Baltimore, Williams & Wilkins.

Strax P (1989) "Control of Breast Cancer through Mass Screening: From Research to Action", Cancer: 63(10) 1881–1887.

Swets JA (1973) "The Relative Operating Characteristic in Psychology", Science: 182 990–1000.

Swets JA & Pickett RM (1982) "Evaluation of Diagnostic Systems: Methods from Signal Detection Theory", New York, Academic Press.

Tabar L & Dean PB (1985) "Teaching Atlas of Mammography", Thieme inc, New York.

Tabar L, Fagerberg CTG, Gad A, Baldetorp L, Holmberg LH, Grontoft O, Ljungquist U, Lundstrom B, Mansson JC, Ekland G, Day NE & Petersson F (1985) "Reduction in mortality from breast cancer after mass screening with mammography:

Randomised trial form the breast cancer screening group of the Swedish national board of health and welfare", Lancet:1, 829–832.

Tabar L, Gad A, Akerlund E, Holmberg L (1983) "Screening for breast cancer in Sweden" in Feig SA, McLelland R (eds) "Breast carcinoma: current diagnosis and treatment", Masson, New York, 315–326.

Thomas JM, Fitzharris BM, Redding WH, Williams JE, Trott PA, Powles TJ, Ford HT & Gazet JC (1978) "Xeromammography and Fine Needle Aspiration Cytology in the Diagnosis of Breast Tumours", Brit Med J: 2, 1139–1141.

Tokanuga M, Norman GE & Assano M (1979) "Malignant tumours among atom bomb survivors, Hiroshima and Nagasaki, 1950–1974", J Nat Cancer Inst: 62, 1347–1359.

Treisman A (1985) "Preattentive Processing in Vision", Computer Vision, Graphics & Image Processing: 31, 156–177.

Triesman A (1988) "Features and Objects: The Fourteenth Annual Bartlett Memorial Lecture", Quart J Exp Psych: 40A, 201–237.

Triesman A & Schmidt H (1982) "Illusory Conjunctions in the Perception of Objects", Cognitive Psych: 14, 107–141.

Tsal Y (1983) "Movements of Attention Across the Visual Field", J Exp Psych: Human Perception & Performance: 9, 523–530.

Turner DA, Alcorn FS, Shorey WD et al (1988) "Carcinoma of the breast: detection with MR versus xeromammograohy", Radiology: 168, 49–58.

Urban J (1976) "Changing Patterns of Breast Cancer", Cancer: 37, 111.

Vernon MD (1971) "The Psychology of Perception", Penguin Books, Middlesex.

Wee WG, Moskowitz M, Chang N–C, Ting Y–C & Pemmerjau S (1975) "Evaluation of Mammographic Calcifications using a Computer Program", Radiology: 116, 717–720.

Welford AT (1976) "Skilled Performance: Perceptual & Motor Skills", US, Scott, Foresman & co.

Yin FF, Giger ML, Vyborny CJ, Doi K & Schmidt RA (1993) "Comparison of Bilateral Subtraction and Single-image Processing Techniques in the Computerised Detection of Mammographic Masses", Invest Radiol: 28(6), 473-481.

Yin FF, Giger ML, Doi K, Metz CE, Vyborny CJ & Schmidt RA (1991) "Computerised detection of masses in digital mammograms: Analysis of bilateral subtraction images", Med Phys: 18(5) 955-963.