Computer Aided Detection of Abnormalities in Mammograms

I W Hutt[1], S M Astley[1] & C R M Boggis[2]

1: Wolfson Image Analysis Unit, Dept of Medical Biophysics, Manchester University, Oxford Rd, Manchester, M13 9PT.

2: Nightingale Breast Screening Unit, Withington Hospital, Nell Lane, Manchester, M20 0PT.

A system that automatically identifies suspicious regions in mammograms could be useful to radiologists by drawing attention to abnormalities that may otherwise have been overlooked. Two detection algorithms are described; one based on the combination of evidence from multiple cue generators and the other based on fuzzy pyramid linking. The latter algorithm proved to be the more effective for locating mammographic abnormalities and was used to generate attention cues, or prompts, for our system. We have performed an experiment in which 100 mammograms were presented to eight radiologists in a manner similar to routine screening practice. These films were presented both with and without prompts and our results demonstrate that the detection performance of the radiologists was significantly improved by prompting.

Introduction

A number of studies have demonstrated the effectiveness of mammography for the detection of breast cancer at an early enough stage to significantly improve the prognosis of the patient [1, 2]. However, the task of interpreting mammographic films can be very difficult, since the radiological signs of early breast cancer can be extremely subtle and are often embedded in complex, highly variable backgrounds.

Using eye-movement data, Kundel and Nodine [3] have studied the errors made by radiologists searching for small nodules in chest x-rays, a task that is analogous to the detection of subtle abnormalities in mammograms. Their studies suggest that up to 40% of film reading errors occur because insufficient attention is directed towards the location of an abnormality. One possible method of reducing errors of this sort is to present radiologists with automatically generated attention cues in conjunction with the films. These cues, or prompts, should act to direct the attention of the radiologists towards potentially suspicious regions of the mammogram that have been identified by computer-based methods.

There is evidence to support the use of prompts as an aid to radiological diagnosis. Treisman [4] suggested that the detection of subtle targets embedded in complex backgrounds requires serial search of the image, and that this process is facilitated by the brief presentation of cues to the target locations prior to presentation of the image. In addition, Chan [5] has conducted a study that demonstrates the use of prompting in mammography for the detection of clustered microcalcifications. She concluded that a

BMVC 1994 doi:10.5244/C.8.34

radiologist working in conjunction with a computer-aided diagnostic system was more effective than either the radiologist or the system working alone.

The aims of our investigation are two-fold: firstly, to study the effects of prompting in as realistic a setting as possible; and secondly, to look at the effects of prompting when there are several types of abnormality present, including some that have not been targeted by the prompt generation system.

Prompt Generation

The first requirement of a prompting system for mammography is some means of automatically detecting potential abnormalities and hence generating the appropriate prompts. We have investigated two methods of prompt generation and conducted a comparative study to assess their accuracy in the detection of clustered microcalcifications, an important sign of early breast cancer. Microcalcifications appear as well-defined, small bright blobs in a variety of shapes and are considered to be clinically significant when they appear in small groups, or clusters.

The first of the algorithms is based on that described by Astley and Taylor [6] and involves the combination of evidence from two cue generators, each of which was selected to respond to a particular property of microcalcifications; the sharp edges, and blob-like appearance. Both of the cue generators are based on mathematical morphology. The first is a morphological inner-edge detector and the second is a top hat transform designed to detect bright peaks of restricted size, in this case up to a diameter of 10 pixels (1 mm). The results of each cue generator are weighted to account for the typical responses of the cue generators to microcalcifications. In order to achieve this, a simple statistical model of typical cue generator responses was derived from a set of 594 known microcalcifications. The weighted cue images are combined by multiplication and subjected to a morphological closing operation to remove very small (100 μm diameter) objects from the combined image. A cluster detection procedure is then used to locate groups of 3 or more potential microcalcifications with nearest-neighbour distances of less than 50 pixels (5 mm).

The second method used for the detection of microcalcifications is a modified version of the fuzzy pyramid linking algorithm described by Brzakovic [7]. The first step in this method is the construction of a gaussian pyramid with the original image as the base. Each level of the pyramid above the base is half the size of the level below it, with the pixel values of the pixels in a level above the base being generated by the application of overlapping 4x4 gaussian masks to the pixels on the level directly below it.

The levels of the pyramid are then linked together by associating every node (pixel) on a level above the base to the 16 nodes on the level below that were used to generate it. The strength of each link is determined by a fuzzy membership function [8] operating on the difference in intensity between the linked nodes. Once all of the links are established, the value of each node is updated by taking a weighted average of the nodes linking with it from below, with the weights corresponding to the appropriate link strengths. 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 uses fixed values for these parameters, while our version fixes α at zero and automatically selects γ to be one standard deviation of the grey level distribution of the original image. This ensures that the full range of possible link strengths is used.

Comparison of Microcalcification Detection Algorithms

To assess the effectiveness of the two detection algorithms, each was tested on the same group of 60 images. The images are 512x512 pixel patches taken from digital mammograms with a spatial resolution of 10 pixels mm^{-1}. Of these 60 patches, 36 contain at least one cluster of microcalcifications and three of these contain two distinct clusters, giving a total of 39 clusters in the data set. The remaining 24 images contain no abnormalities. The locations of the clusters were identified by a radiologist using both the digital images and the original films. In the case of the first algorithm, test films were also required for training the system using a 'leave-one-out' approach.

In order to generate points for receiver operating characteristic (ROC) curves, 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 free-response ROC curves shown in figure 1. A statistical analysis of the performance of each algorithm in discriminating between normal and abnormal films revealed that the performance of the pyramid algorithm was significantly better than that of the cue combination method ($t_{obs} = 4.17$, $p < 0.005$).

In this test the performance of the fuzzy pyramid algorithm exceeded that of the morphological system, with the 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 is operating at a level of accuracy 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 positive clusters generated at these operating points were approximately 2.5 per image and 6.1 per image respectively. Previous research has suggested that the benefit of prompts as aids to the radiologist is diminished and may be lost altogether as the false-positive rate of the prompt generation system increases [9]. Even the rate of 2.5 false-positives per image achieved by the fuzzy pyramid system may be too high for the prompts to be useful in clinical practice.

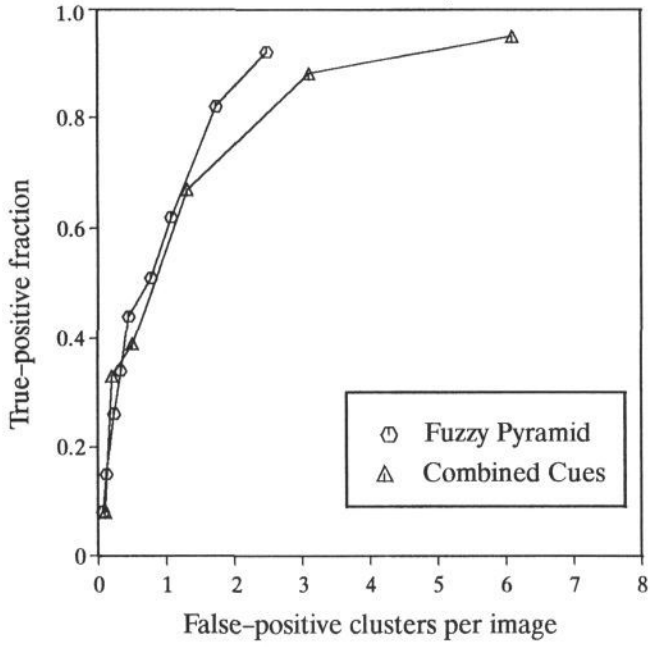

Figure 1: FROC curves showing detection performance of algorithms

At present, the final stage of processing in the fuzzy pyramid algorithm consists of simple tests to determine the sizes of potential microcalcifications in the segmented image and to decide whether or not they represent a cluster. By introducing some more sophisticated feature testing, it may be possible to improve the specificity of the system. For example, it might be useful to determine the locations of any detected potential microcalcifications in the segmented image and examine the properties of these locations in the original image with a view to rejecting any that represent clearly normal tissue.

Prompting Experiments

One of our principal aims is to study the effects of prompting in a setting and task as similar as possible to clinical film-reading practice. For this reason, all of the participating radiologists read the films in their own reporting rooms, using the viewing equipment used during screening sessions.

In these experiments, the prompt generation system was based on the fuzzy pyramid algorithm and was used to target both clustered microcalcifications and tumours. For the detection of tumours, the thresholding procedure was modified to look for strong links associated with the smooth internal structure of such lesions and the cluster detection procedure was replaced by a size test that responded to tumour-sized objects (> 1 cm diameter) in the segmented image.

A total of 100 pairs of mammograms were used in the study, all of which were obtained by routine screening. Each mammogram pair consisted of the medio–lateral views of the left and right breasts of a single patient. Half of the films (50 pairs) were normal, while the remainder were divided equally among five groups; single microcalcification clusters, single well–defined lesions, single untargeted abnormalities (either spiculated lesions or architectural distortion), multiple targeted abnormalities and multiple abnormalities including at least one untargeted abnormality. In each of these five groups, half of the cases were malignant and half were benign.

Each film was digitised with a sampling rate of 100 μm per pixel and an 8–bit grey resolution. Each digitised image was then divided into five overlapping 1024x1024 pixel regions for processing by the prompt generation system.

The prompt generation system produced approximately one false–positive per region, which resulted in an average of about 10 invalid prompts per pair of films. This was clearly an unacceptable false–positive rate, so the number of false–positives was reduced by re–processing the films with strict criteria for the prompt generation algorithm, so that very few prompts were generated. The false positives from this second round of processing were combined with the true positives from the original processing, to simulate a system of greater specificity than was actually the case.

86% of the microcalcification clusters and 67% of the tumors were prompted. Although the prompt generation algorithm was not designed to detect the untargeted abnormalities, 35% of these were actually found and prompted. The (simulated) false–positive rate across all of the images was 1.1 invalid prompts per film pair.

Each of the digitised film pairs was printed out on a laser printer to produce two low resolution hard copies of the mammograms. One of these copies showed the films in the unprompted form, while the other showed the prompted version, with the prompts superimposed as dark circles on the images.

Eight consultant radiologists, all with extensive experience in reading mammograms, took part in the study. 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 prompted or unprompted version, depending on the experimental condition. For each session, all 100 film pairs were loaded onto the viewer in a random order.

At the beginning of each session, the participating radiologist was provided with written instructions on the task and required responses. Each participant was told to expect around half of the films to contain abnormalities. When prompting was used, the radiologist was also told that the prompt generation algorithm was specifically targeting microcalcifications and tumours.

Each of the participants was presented with each film in both the prompted and control conditions, with the presentations divided between two sessions of 100 films each. In each case the two sessions took place on different days and the two versions of any given film were always presented in separate sessions.

In addition to serving as media for the prompt information, the hardcopy mammograms also acted as response forms. Each hardcopy consisted of a film number for reference, the copy mammograms with or without prompts superimposed and a six point rating scale as follows:

0: Normal
1: Benign
2: Probably benign
3: Uncertain
4: Probably malignant
5: Malignant

The points of the rating scale correspond to those generally used to rate films in screening centres with the exception of the '0: Normal' point. In screening, normal 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 a rating other than zero was given to the film, the radiologist was also requested to mark on the hard copy the location of any detected abnormalities.

The truth data for these experiments were provided by a consultant radiologist who had access to both the original films and the patient records. The consultant marked the locations of all abnormalities on acetate overlays which were registered with the hard copy images. The responses of the participating radiologists were compared with these annotations to calculate the numbers of true-positive and false-positive responses given at each point on the rating scale. An abnormality was considered to have been correctly located if the location marked by the subject fell within 1 cm of the known centre of the lesion.

Results and Discussion

The performance of the radiologists in each experimental condition was assessed by producing a free response operating characteristic (FROC) curve for each subject with every point on the rating scale representing a different level of response bias [10]. The results for individual subjects were then pooled by averaging the numbers of true-positive and false-positive responses at each criterion level. Figure 2 shows the pooled FROC curves for both the prompted and control condition.

Conventional ROC analysis was also used to examine the results. In order to obtain the required classification data, the highest rating for each film was used as the criterion level and the scale points of '0: Normal' and '1: Benign' were collapsed into a single category. Again, an ROC curve was obtained for each participating radiologist and the results were pooled by averaging the values for the true-positive and false-positive fractions at every level. Figure 3 shows the pooled ROC curves for each condition. The solid curves in figure 3 represent best-fit curves obtained using the ROCFIT program [11].

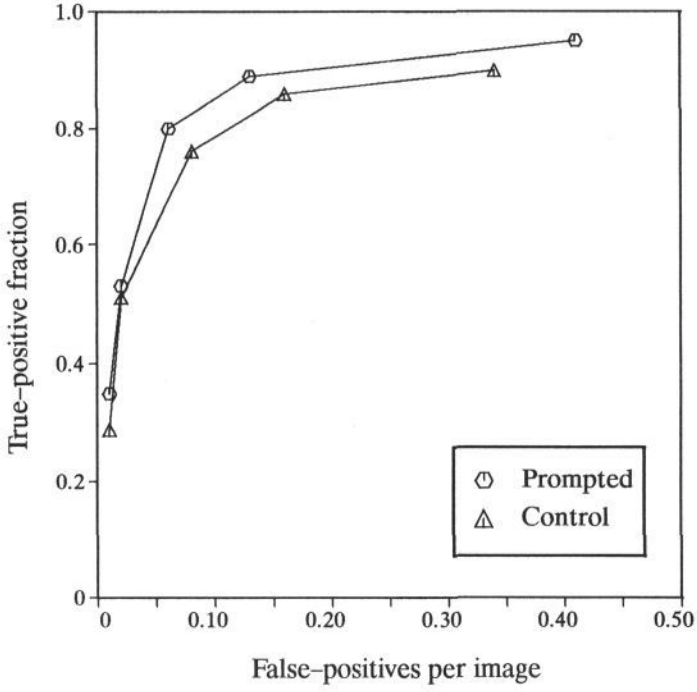

Figure 2: FROC curves showing performance in each condition

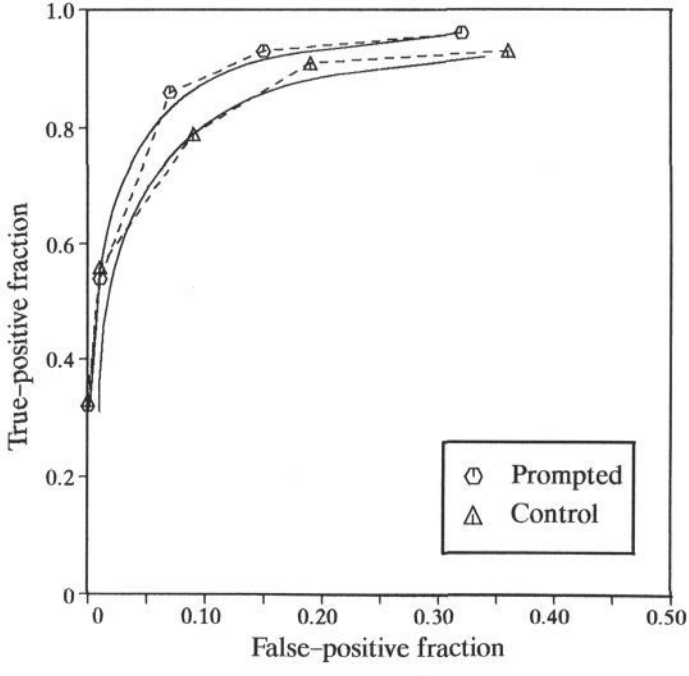

Figure 3: ROC curves showing performance in each condition

Values for the sensitivity index d_a [10] were calculated for each radiologist under each experimental condition. Analysis of these measures reveals that the overall detection accuracy of the radiologists was significantly higher in the prompted condition than in the unprompted condition, ($t_{obs} = 4.13$, $p < 0.005$). Table 1 shows the values of d_a for each radiologist in both conditions.

As can be seen from table 1, the improvement in detection sensitivity that occurred with prompting was remarkably consistent. Each subject demonstrated a higher detection sensitivity in the prompted condition than in the control.

Table 1: Values of d_a for each radiologist

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

All of the participating radiologists demonstrated a substantially higher detection performance than 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. Thus the results of this study support 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 [5]. The reason for this increase in detection performance is that, with the aid of prompts, the radiologists correctly located abnormalities that were otherwise missed. Of the 600 abnormalities presented in the study (75 abnormalities to each of 8 radiologists), 37 were only identified when they were prompted.

In four cases in which abnormalities were detected in the control condition but missed in the prompted condition and in all of these cases the abnormalities had been missed by the prompt generation system. It appears that in these four instances, prompting actually caused the radiologists to miss abnormalities that they would otherwise have detected.

A possible explanation for this effect may lie in the observation that apart from being missed by the prompt generation system, three of these four cases share another common factor; the presence of at least one false-positive in the image. This combination of a false-negative and a false-positive error on the same film leads to a situation in which a prompt, rather than directing the attention of the observer towards an abnormality, directs attention away from it and towards an area of normal tissue. There were 16 instances in the 100 films where this type of combination error occurred, a total of 128 cases among the eight subjects, and in only in three instances did invalid prompting lead to abnormalities being missed. Therefore, if combination errors were

responsible for the effect, there must be some additional conditions that were not present in all of the cases. Although this does not appear to be a particularly significant effect, it could indicate a serious drawback for the use of prompting and therefore warrants further investigation.

A related problem is the extent to which false-positive prompts may lead the radiologist to make false-positive responses that they would not otherwise have made. Table 2 shows the number of false-positive judgement made by each radiologist in each condition. Although there is a general trend towards an increase in the number of false-positives in the prompted condition, this increase falls short of statistical significance ($t_{obs} = 1.36$).

Table 2: Numbers of false-positive responses in each condition

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

Although the results suggest that in this case invalid prompts have not had a significant detrimental effect on the detection performance of the radiologists, it should be noted that the results of the prompt generation system were modified by reducing the false-positive rate and that if all of the original invalid prompts had been included the effectiveness of the prompts may have been reduced.

It is not our intention to provide any actual figures for the level of improvement in performance that might be observed if prompting were to be introduced into routine mammographic screening. Far more extensive testing would be required before such a move could be contemplated. However, this experiment has demonstrated that prompting can be an effective aid to radiologists in the detection of a range of types of abnormalities in mammograms in an environment that approaches clinical screening practice.

References

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

[2] 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 from the breast cancer screening group of the Swedish national board of health and welfare", *Lancet*, 1: 829–832.

[3] Kundel HL & Nodine CF (1978) "Studies of eye movements and visual search in radiology" in Senders JAW, Fisher D & Monty R (eds) "Eye movements and the higher psychological functions", Hillsdale NJ, LEA.

[4] Treisman A (1985) "Preattentive processing in vision", *CVGIP*, 31: 156–177.

[5] 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.

[6] Astley SM & Taylor CJ (1990) "Combining Cues for Mammographic Abnormalities", *Proc BMVC 1990*, 253–8.

[7] Brzakovic D, Luo XM & Brzakovic P (1990) "An approach to automated detection of tumors in mammograms", *IEEE Trans – Medical Imaging*, 9/3: 233–41.

[8] Zadeh LA (1965) "Fuzzy Sets", *Information and Control* 8: 338–53.

[9] Hutt IW (1992) "The effects of prompting on the detection of clustered microcalcifications in digital mammograms", MSc Thesis, Manchester University.

[10] Macmillan NA & Creelman CD (1991) "Detection theory: A user's guide", Cambridge, CUP.

[11] Metz CE (1989) "Some practical issues in experimental design and data analysis in radiological ROC studies" , Invest Radiol, 24: 234–245.