Combining Cues for Mammographic Abnormalities

Susan M. Astley and Christopher J. Taylor

Department of Medical Biophysics
University of Manchester,
Oxford Road,
Manchester M13 9PT

Screening for breast cancer involves searching for subtle abnormalities in a large number of complex images, a task for which the specificity of human interpreters is known to be poor. We are aiming to improve screening performance by providing radiologists with machine assistance in the detection of clinically significant features. The first stage of the detection process is the generation of a set of cues to indicate potential abnormalities. We generally select a cue method for a particular task because it responds to a known property of the target. However, cue generators also respond to non-targets which share that target property. By combining evidence from a range of cues associated with different target properties we can increase the specificity of detection in noisy or cluttered images. We have performed experiments which demonstrate this. Two cue generators were applied to a set of 20 digitised image patches. On- and off-target distributions were collected for each image and accumulated across the data set on a leave-one-out basis. Each cue image was then transformed into a log-likelihood image, enabling evidence from the different cue generators to be combined simply by image addition. Results of an evaluation of single and combined cue methods are presented.

Breast cancer is a potentially fatal disease that affects about one in twelve women at some time in their lives. A national breast screening programme has recently been instituted with the aim of enabling effective treatment of the disease by detecting it at an early stage in asymptomatic women; this programme is expected to generate over 1.5 million breast X-rays (mammograms) per year. Mammographic images are highly variable and often complex. Breast cancer screening involves systematically searching these images for abnormalities. Two important mammographic signs of early breast cancer are clusters of microcalcifications, which appear as groups of very small, sharp-edged blobs brighter than their background, and spiculated lesions, which are characterised by radiating linear structure. Such signs may be subtle and high intra- and inter-observer variabilities have been reported [1].

Machine assistance may prove valuable either to pre-select equivocal and abnormal films for subsequent detailed analysis or to draw the radiologist's attention to suspected abnormalities. Previous attempts to automate the detection of microcalcifications have used sequences of progressively more sophisticated methods to refine a set of candidates e.g. [2], though clinically acceptable error rates have not yet been achieved.

The first stage of the detection process is the generation of a set of cues; these indicate which regions of an image are of further interest. Many techniques have been developed for extracting simple properties, such as regions, edges, linear structures and corners [e.g. 3,4,5] from images. For any given application, we generally select a cue generation method because it will respond selectively to some particular characteristic of the target we are trying to detect. It will, however, also respond to non-targets in the background which possess that characteristic, and consequently methods often perform poorly in complex or noisy images. We can improve the specificity and robustness of the target-detection process by combining information from a number of independent image cues, each responding to a different characteristic of the target. On-target responses in the different cue images will be mutually supportive, whereas off-target responses may conflict. Little work has been done on the systematic combination of image cues except at very high levels, for example stereo and motion, shape and stereo etc [6]. The approach is related closely to studies of natural (human) image analysis which suggest the involvement of multiple cues [7].

CUES

As our primary objective is to demonstrate the feasibility of improving detection performance by combining cues, we have used two readily available cue generators in our initial experiments. These were chosen to respond selectively to different properties of microcalcifications. Both methods are based on mathematical morphology [4] and use structuring elements which approximate to a uniform disc. The first is the morphological top hat transformation, formed by subtracting an opened image from the original, which we use to preferentially enhance topographical peaks of restricted size. The second is a simple morphological edge detector, which we expect to respond to the sharp edges of the microcalcifications. Since we intend to combine the outputs of these two methods, we must ensure spatial correspondence of the edge responses with the peak responses. We therefore use a detector of the same

BMVC 1990 doi:10.5244/C.4.45

edges of bright objects, produced by subtracting an eroded image from the original.

CUE COMBINATION

We are seeking to combine, in a principled way, cue generator responses, which are generally scaled in an arbitrary fashion. Our approach is to estimate the probability of a true response at each image point for each cue independently and to combine these probabilities using Bayesian statistics [8]. We can do this by gathering data from a similarly-scaled set of training images in which the positions of the targets have been identified. Using Bayes rule we can write:

$$P(\text{lesion}|x) = P(x|\text{lesion})\frac{P(\text{lesion})}{P(x)} \qquad \text{and}$$

$$P(\overline{\text{lesion}}|x) = P(x|\overline{\text{lesion}})\frac{P(\overline{\text{lesion}})}{P(x)}$$

where x represents a cue generator response and P(x) the probability of getting that response. We can obtain P(x|lesion) by examining cue generator responses to known abnormalities in the training set, and P(x|$\overline{\text{lesion}}$) by looking at genuine background responses. It is less straightforward to establish an appropriate estimate of P(x), since training data selected for our experiments are biased to provide more on-target information than one would obtain from a randomly selected set of screening films. We can, however, eliminate P(x) by dividing the above equations to obtain an expression for the posterior odds, O(lesion|x), or strength of belief that a cue generator response x represents a lesion:

$$O(\text{lesion}|x) = \frac{P(\text{lesion}|x)}{P(\overline{\text{lesion}}|x)} = \frac{P(x|\text{lesion})}{P(x|\overline{\text{lesion}})}\frac{P(\text{lesion})}{P(\overline{\text{lesion}})}$$

That is,

$$O(\text{lesion}|x) = L(x|\text{lesion})O(\text{lesion})$$

since $\dfrac{P(x|\text{lesion})}{P(x|\overline{\text{lesion}})}$ is the likelihood ratio,

and $\dfrac{P(\text{lesion})}{P(\overline{\text{lesion}})}$ is the prior odds of finding a lesion.

The likelihood ratio can be calculated from training data, and enables us to transform cue images into a form amenable to combination. The prior odds, an estimate of which can be derived from clinical information, acts merely as a threshold.

Suppose we have N cue generators, and wish to combine the (possibly conflicting) evidence they provide. In this case,

$$O(\text{lesion}|x^1, x^2, \ldots, x^N) = L(x^1, x^2, \ldots, x^N|\text{lesion})O(\text{lesion})$$

where x^k represents the response of the k-th cue generator. Assuming independence of the cues,

$$P(x^1, x^2, \ldots, x^N|\text{lesion}) = \prod_{k=1}^{N} P(x^k|\text{lesion}) \qquad \text{and}$$

$$P(x^1, x^2, \ldots, x^N|\overline{\text{lesion}}) = \prod_{k=1}^{N} P(x^k|\overline{\text{lesion}}) \ .$$

We can thus write

$$O(\text{lesion}|x^1, x^2, \ldots, x^N) = O(\text{lesion})\prod_{k=1}^{N} L(x^k|\text{lesion})$$

i.e. we can combine evidence from our set of cue generators by computing the product of likelihoods. The method allows the assimilation of evidence from additional cue generators without having to recompute, since

$$P(\text{lesion}|x_N, x) = P(\text{lesion}|x_N)\frac{P(x|x_N, \text{lesion})}{P(x|x_N)}$$

where x_N are a set of N cue generator responses, and x is the response of a new cue generator. If we again assume independence, we see that we can easily include the new response, since

$$P(x|x_N, \text{lesion}) = P(x|\text{lesion}) \ ,$$

$$P(x|x_N, \overline{\text{lesion}}) = P(x|\overline{\text{lesion}})$$

and hence

$$O(x|x_{N+1}) = O(\text{lesion}|x_N)L(x|\text{lesion})$$

EXPERIMENTAL METHOD

Our aim was to demonstrate that this method of cue combination could indeed lead to greater specificity in detecting mammographic abnormalities. To this end, we have carried out an investigation comparing the

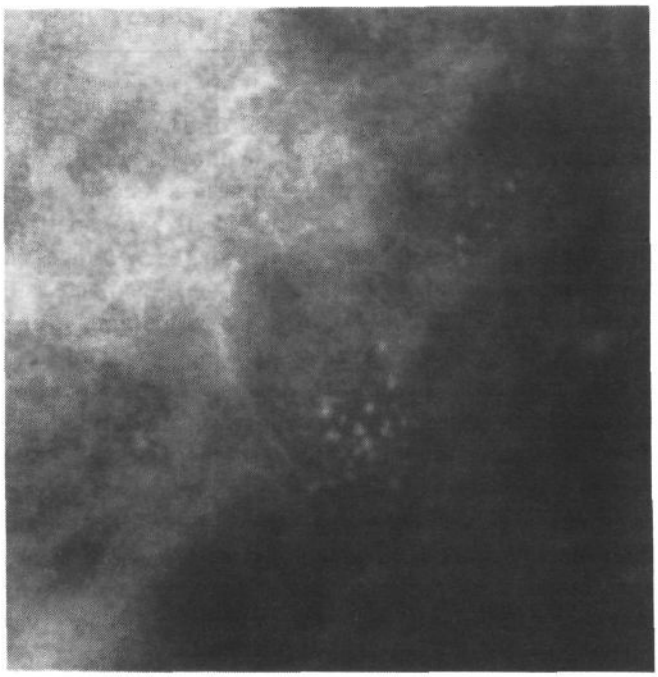

Figure 1. Digitised 2.5 cm square patch of mammogram showing malignant microcalcifications (small, bright blobs).

performance of two cue generation methods with their combined performance.

A chronological sequence of twenty screening mammograms was obtained from the Manchester Breast Screening Service. These were taken consecutively from the set of those showing biopsy-proven cancer, for which the biopsy was performed on account of the presence of microcalcification. A consultant radiologist selected a 2.5cm by 2.5cm patch of each mammogram, including in each patch at least one cluster of microcalcifications; these patches were digitised at 20 pixels per mm (e.g. figure 1). Two radiologists attempted to identify the location of the centre of every microcalcification falling within the selected patches in the original film images. Microcalcifications were marked by the radiologists with a fine pen on an acetate overlay. Each acetate was then registered with the corresponding original image and digitised. Pen-marks in the digitised acetate images were detected by thresholding and reduced to a uniform size. They were then reviewed - and in some cases edited - by the radiologists, using both film and magnified digital images to assist in checking their validity. A total of over 900 individual microcalcifications were identified in all. The set of identified abnormalities corresponding to the patch shown in figure 1 is displayed below in figure 2.

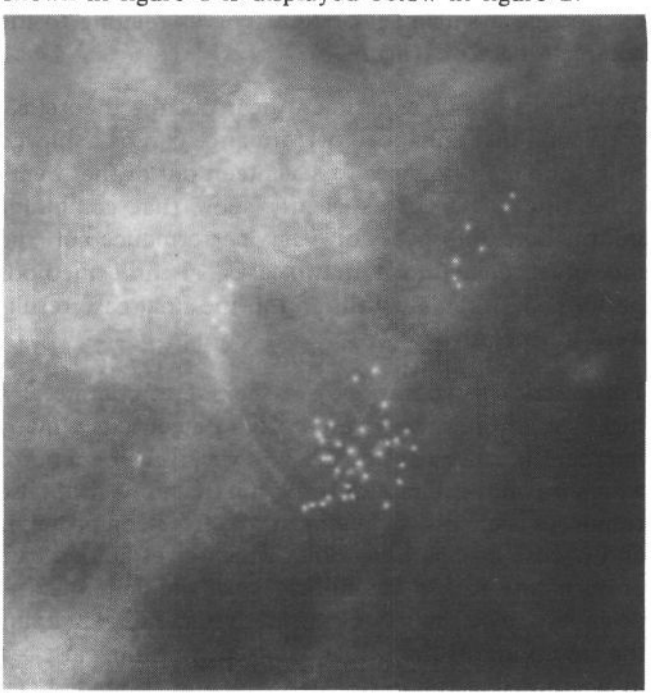

Figure 2. The example image from figure 1 with microcalcifications identified by a consultant radiologist highlighted.

As it was our intention to gather on- and off-target cue generator responses to the identified microcalcifications, we created a region of interest (ROI) around each of the marked microcalcification positions. These were defined such that each ROI would contain responses to at least one microcalcification, and that no response to a genuine microcalcification would appear outside the set of ROIs. Our initial experiments demonstrated that significant loss in performance could be caused by lack of care in collecting the training data, and that the definition of ROIs was problematic because of the competing requirements to include a reasonably large area around each microcalcification whilst not merging ROIs for clustered microcalcifications. Our current method for defining the ROIs is as follows: we first associate each pixel with the nearest microcalcification marker by applying a two-pass distance transform [9] to the image of microcalcification markers to produce a a Voronoi diagram [10]. Voronoi polygons associated with markers close to the edge of an image are excluded from all subsequent analyses. We next construct a set of discs, each centred on a microcalcification marker. The disc radius is sufficiently large to encompass all cue generator responses to the underlying microcalcification; a radius of 16 pixels was used for these experiments. Each ROI is defined by the intersection of a Voronoi polygon and its associated disc. Off-target responses for an image patch are gathered from the region which is the inverse of the set of ROIs for that patch.

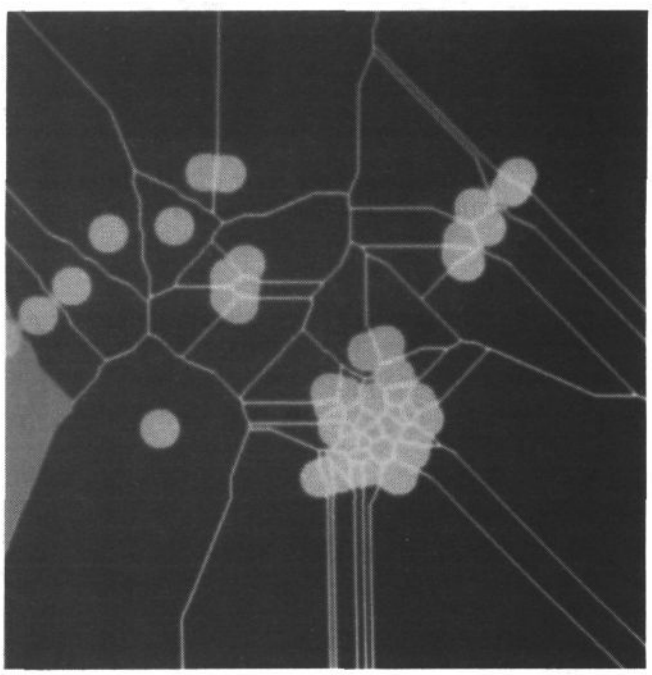

Figure 3. The Voronoi diagram (white), intersections of Voronoi polygons with discs centred on microcalcification markers (light grey), and the border region excluded from analysis (dark grey).

Two cue generators, selected to respond to different target features, were applied to each image. These were the morphological top hat transformation, with a structuring element of diameter equivalent to 0.6mm in the original film mammogram, and a morphological "inner edge" detector (figure 4, (a) and (b)). For each cue generator, the maximum value within an ROI was taken to represent the on-target response. All responses not corresponding to a region of interest (with the exception of those falling within any Voronoi polygon associated with a disc near the border of the patch) were

classified as off-target responses. On- and off-target distributions of the responses of each cue generator were collected for all the image patches, and accumulated across the data set on a leave-one-out basis. Since these represent P(x|lesion) and P(x|$\overline{\text{lesion}}$), we were able to calculate the likelihood ratio and hence transform cue images into lesion likelihood images.

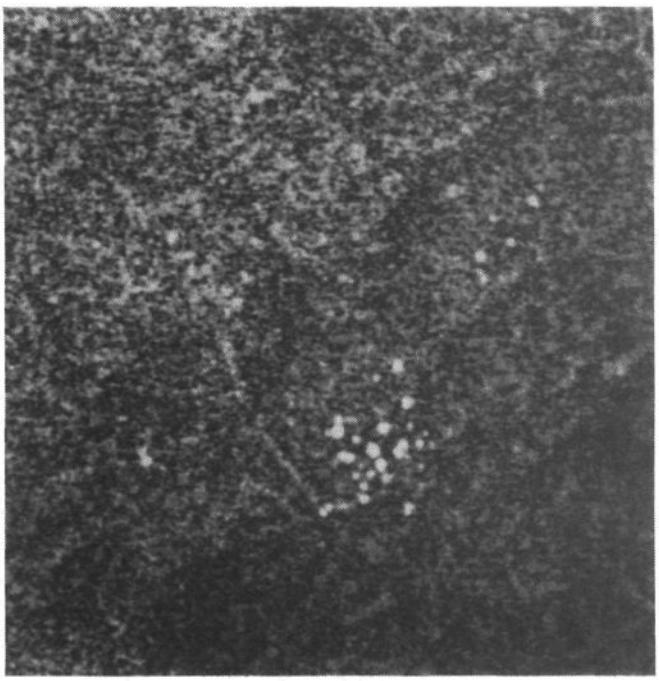

Figure 4(a). Output of the morphological top hat transformation applied to the image shown in figure 1.

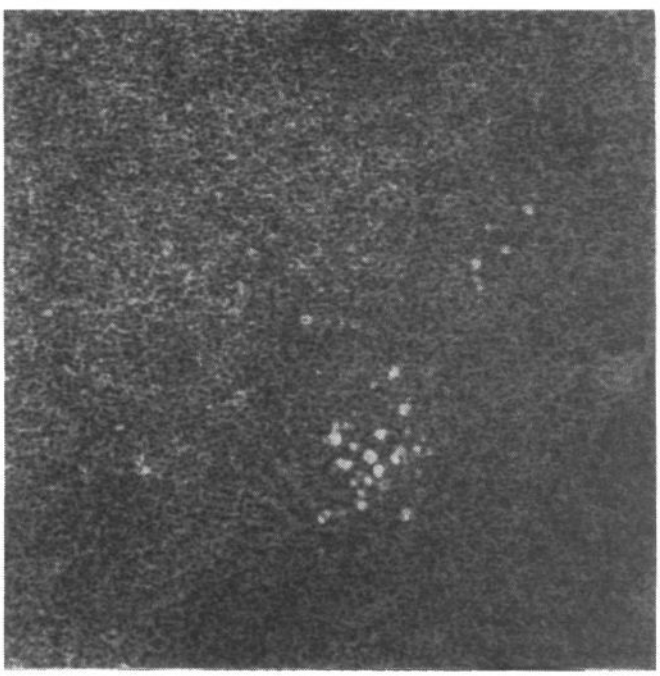

Figure 4(b). Morphological edge detector output for the same example image.

In practice, the dynamic range of the data is such that it is more convenient to a compute log likelihood images for each cue generator (figure 4 (c) and (d)). Each image was transformed using data collected from a training set comprising the other nineteen images. Evidence from the two cue methods was then combined by adding their log likelihood images (figure 5), and the performance of the combined cue method was compared with that of the single cue methods.

RESULTS

To assess the performance of the three methods, on- and off-target data were collected from the two log likelihood images and from the combined log likelihood image. These distributions were all scaled to unit area. True and false positive fractions were computed for each method by summing responses above the same threshold on the on- and off-target distributions.These were computed across the range of thresholds and plotted as receiver operating characteristic (ROC) curves [11] in which the percentage of true positive responses (i.e. the percentage of genuine abnormalities detected) is shown against the number of false positive responses per $100cm^2$, an area comparable with the area of interest in a typical mammographic film. This method of expressing the results in terms of a specified area will allow direct comparison of our results with those of other authors. ROC curves for the three cue methods were computed for all twenty images. Data from all the log likelihood images were also accumulated to produce a single set of ROC curves relating to the complete data set (figure 6).

The results of our experiment bear out the theoretical prediction that enhanced performance can be achieved by combining evidence. In seventeen out of the twenty cases, the ROC curves clearly show that combining information improves on the performance of the individual cue methods. In two cases, all three methods performed extremely well, and in the final case results were inconclusive.

For practical purposes, we are seeking to operate at a high true positive rate, to detect as many microcalcifications as possible. However, the benefits of a high true positive rate will be lost if the number of false responses is excessive. We must therefore compromise, and choose an operating point which gives a good true positive rate with a reasonably small number of false positives. Taking this into consideration, and examining the results across the whole data set, we can clearly see in figure 6 that the performance of the combined cue method is significantly better than that of either of the single cue methods.

DISCUSSION

In developing machine-assisted screening methods we must strive for methods which are both sensitive and specific, since the consequences of false negative mammographic interpretation (missed cancers) and false positive interpretation (traumatic, expensive investigation) are both serious. Cue combination is one method by which we can improve on the performance of

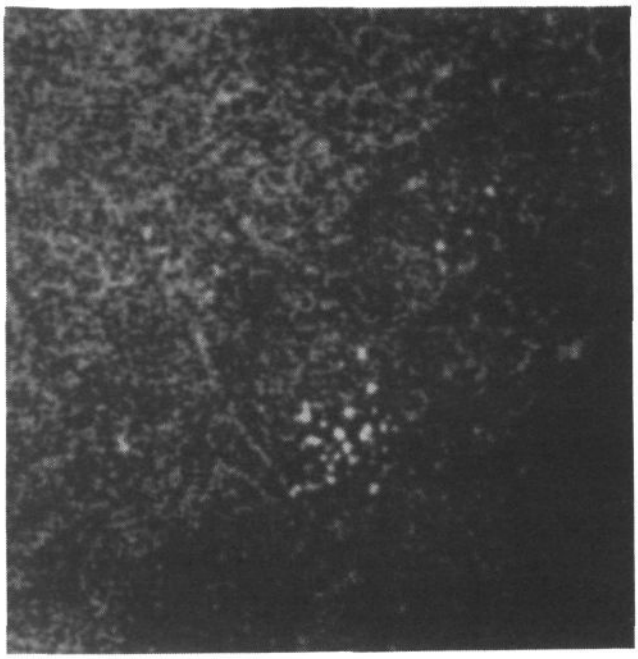

Figure 4(c). The top hat transformation image shown in figure 4(a) converted to display the log likelihood of microcalcification being present based on on- and off-target distributions gathered from nineteen other top hat images.

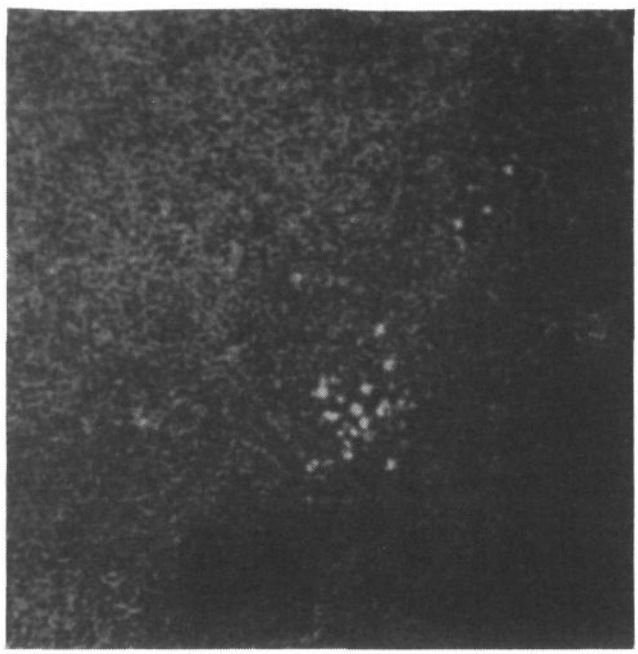

Figure 4(d). The morphological edge image shown in figure 4(b) converted to display the log likelihood of microcalcification being present based on on- and off-target distributions gathered from nineteen other morphological edge images.

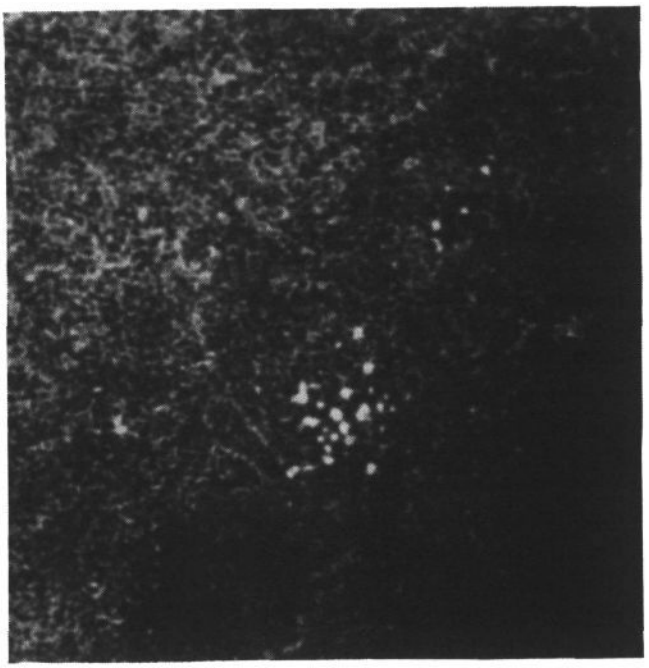

Figure 5. The log likelihood image showing combined evidence from the morphological edge detector and the top hat transformation.

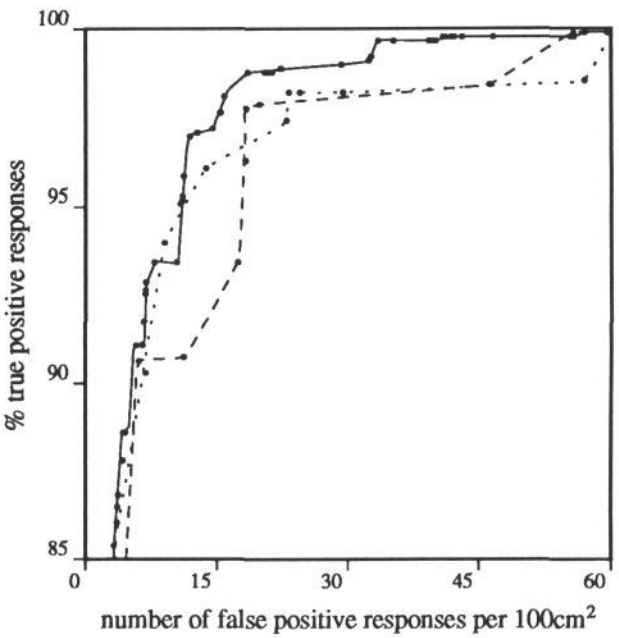

Figure 6. Receiver Operating Characteristic curves for edge cues (dashed line), top hat transformation cues (dotted line) and combined cues (solid line) using data gathered from twenty images. The percentage of true positive responses (detected microcalcifications) is plotted against the number of false positive responses per 100cm².

individual cue generation methods to detect mammographic abnormalities.

There are, however, a number of practical problems associated with this approach to the problem. Firstly, there is a genuine difficulty in defining true positive responses in the training data set. Radiologists rarely, in practice, search mammograms (let alone magnified digital images) for *all* microcalcifications. The process of identifying and marking microcalcifications is tedious, subjective and error-prone; only after a process of re-examination and refinement was a reasonably

satisfactory solution achieved. Mammograms are projection images, so individual particles in clusters may appear to overlap, touch or be closely adjacent. Our method of defining ROIs, whilst generally satisfactory, does allow some transfer of responses between very closely adjacent microcalcifications. We do not, however, believe that this has significantly affected our results. Secondly, the selection of the maximum response within an ROI introduces a bias towards stronger responses. Thirdly, the data set itself was, of necessity, biased to show a much larger number of abnormalities than one would find in an average screening selection.

We are currently investigating alternative and additional microcalcification cues. In particular, the examples in figure 4 illustrate the necessity of the inclusion of a shape-selective operator to eliminate responses to streaks of normal breast tissue which are, at present, detected both by the top hat transformation and by the morphological edge detector. One of the advantages of the method we have described is the ease with which such additional cue information can be assimilated. It is important to note that this depends on the assumption that the cue generators are independent. We could alternatively take any dependency into account, but this would complicate the assimilation of new evidence.

We have also been investigating methods of cue generation for spiculated lesions; these appear in mammograms as foci of radiating linear structure. We are using a set of measures in Hough transform space which characterise different properties of star-shaped patterns of lines [12]. Each image point is considered as a potential focus for a lesion and the measures characterise the spatial organisation of linear structures about the point in terms of degree of focus, spread of directions etc.. These measures are currently stored in the form of a set of cue maps, which we intend to combine using an adaptation of the method presented in this paper.

ACKNOWLEDGEMENTS

We would like to thank the staff of the Manchester Breast Screening Centre for assistance and advice during the course of the project. In particular, we would like to thank Dr. David Asbury, Clinical Director of the Centre, and Drs. Caroline Boggis and Mary Wilson for their enthusiastic support for this work, for defining our data set and for identifying microcalcifications. We are grateful to the North Western Regional Health Authority for funding, and to IBM UK Scientific Centre for providing a Fellowship for SMA.

REFERENCES

1. **Gale, A.G., Walker, G.E., Roebuck, E.J. & Worthington, B.S.** "The quest for accuracy, consistency and uniformity of performance in mammographic screening: The systematic imperative" *Br J Radiol*, 62 (1989), S10.
2. **Chan, H-P., Doi, K., Vyborny, C.J., Lam, K-L, & Schmidt, R.A.** "Computer-aided detection of microcalcifications in mammograms: Methodology and preliminary clinical study" *Inv Radiol*, 23 (1988), pp 664-671.
3. **Marr, D.** *Vision*, W. H. Freeman, New York (1982).
4. **Serra, J.** *Image analysis and mathematical morphology*. Academic Press, London (1982).
5. **Kitchen, L. & Rosenfeld, A.** "Grey-level corner detection" *Patt. Rec. Lett.* (1982), 1, pp 95-102.
6. **Aloimonos, J.Y. & Brown, C.M.** "Robust computation of intrinsic images from multiple cues" *Advances in Computer Vision* Vol 1 (1988) LEA, New Jersey.
7. **Treisman, A.** "Preattentive processing in vision" *Comp. Vis. Image Proc.* (1985), 31, pp 156-177.
8. **Pearl, J.** *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Mateo (1988).
9. **Rutovitz, D.** "Expanding picture components to natural density boundaries by propagation methods. The notions of fall-set and fall-distance" *Proc 4th Int. Joint Conference on Pattern Recognition*, Kyoto (1978) pp657-663
10. **Preparata, F. P. & Shamos** *Computational Geometry* Springer Verlag (1985) ch 5
11. **Green, D.M. & Swets, J.A.** *Signal detection theory and psychophysics*, Wiley, New York (1966)
12. **Ellison, T.P.** *Detection of Stellate Lesions in Mammographic Images*. MSc Thesis, University of Manchester (1988).