Archived at the Flinders Academic Commons:
http://dspace.flinders.edu.au/dspace/

‘This is the peer reviewed version of the following article:

which has been published in final form at

http://crpit.com/abstracts/CRPITV97Maeder.html

Copyright © 2009, Australian Computer Society, Inc. This paper appeared at the Australasian Workshop on Health Informatics and Knowledge Management (HIKM 2009), Wellington, New Zealand. Conferences in Research and Practice in Information Technology (CRPIT), Vol. 97. J. Warren, Ed. Reproduction for academic, not-for profit purposes permitted provided this text is included.
Assessing Viewing Pattern Consistency in Mammogram Readers

Anthony Maeder* and Clinton Fookes+
*School of Computing and Mathematics
University of Western Sydney, Locked Bag 1797 Penrith South DC, NSW 1797
+ Image & Video Research Laboratory
Queensland University of Technology, GPO Box 2434 Brisbane, Qld 4001
a.maeder@uws.edu.au  c.fookes@qut.edu.au

Abstract
Breast cancer screening programs typically require very large volumes of x-ray images (mammograms) to be viewed by highly experienced human readers. The readers can recognise a wide range of different visible features indicative of clinically abnormal situations, which they use as a basis to generate a report on their findings. Errors in reporting can occur if the readers fail to identify a particular feature of interest for further visual inspection during the viewing process. This risk is typically reduced by training readers to follow a particular viewing path through an image, which they should be able to apply consistently. Knowledge of the extent of consistency in this viewing behaviour within and between viewers would inform the development of an automated checking approach, based on monitoring of viewer visual attention. This paper presents an analysis of some reader viewing pattern profiles obtained using eye tracking with an infra red computer vision system, as a basis for developing a suitable consistency assessment model.

Keywords: medical image, eye tracking, visual attention.

1 Introduction

Many developed countries have instituted national breast cancer screening programs based on 2D X-ray imaging of the compressed breast, typically available every two years for women between about 40 and 60 years of age. This approach has been shown to be highly cost-effective and efficient at detecting many cancers. In Australia, it is estimated that approximately 1 million women are screened each year, which involves acquisition of at least 4 high resolution X-ray images (or mammograms) per subject. These images are subsequently scrutinised independently by at least 2 and sometimes up to 4 highly skilled human readers, who have been trained specifically for this task and are subject to ongoing quality assurance or training processes to maintain their performance level.

When a reader decides that abnormalities which could indicate the formation of cancers are visible in the mammograms, he or she reports these accordingly and the subject is recalled for further clinical assessment.

Readers are trained to follow a particular viewing path through the image, associated with landmark anatomical sites (e.g. armpit, nipple, ductal region, chest wall). Usually the mammograms of both breasts are viewed next to each other, and the reader inspects the same anatomical locations in each image together during the viewing sequence so that asymmetry of features can be used to help detect abnormalities. Readers can recognise a wide range of different visible features indicative of abnormal situations, such as internal lesions in the breast tissue, or microcalcifications in the breast vessels. Whether such features indicate abnormalities or not to the reader, is highly dependent on the “context” provided by the surrounding tissue and the overall breast morphology (e.g. density, texture). Readers may also have access to prior mammograms or previous assessment images from the last screening cycle, to allow them to make visual comparisons and detect changes in tissue characteristics. Consequently, the reader must perform a complex multiple-matching pattern recognition task, based on their expertise acquired from viewing many thousands of mammograms including carefully selected educational examples. Readers are required to find all abnormalities when reporting an image, so they must be wary of a “satisfaction of search” effect which reduces their performance once one feature of interest has been established as yielding a positive result.

False positive and false negative rates are both important measures for assessing reader performance. False positives cause increased costs and patient inconvenience or discomfort due to the additional assessments, while false negatives prevent cancers from being treated early when there is the greatest chance of success. Viewer behaviour models based on visual saliency (e.g. Itti and Koch 2000) have suggested that readers can be triggered to notice features of interest by both overt gaze fixations and covert peripheral visual attention attraction. The consistency of the viewing path is therefore as important for positioning the gaze of the viewer “in the vicinity” of the feature, as it is for directing the gaze directly at the feature. However, there is little understanding of the mechanisms which cause failure to identify features of interest, once common influences like stress, fatigue and distraction have been excluded. Other factors influencing reader performance are the variations in image appearance due to uncontrolled ambient viewing conditions and different intensity and magnification settings, which readers currently address by using a familiar viewing environment and some optical or mechanical aids attached to the viewing station (e.g. magnifiers, tubes, hoods).
The current international trend to replace film-based mammograms with digital images offers some opportunity to improve both viewer performance and our understanding of the various above factors which influence viewer performance. There is considerable evidence to suggest that use of digital mammograms can increase both the true positive and false positive rates (Pisano et al 2005; Hambly et al 2008). One reason for this effect is that digital images allow standardisation or normalisation of image display characteristics, which can help to remove perturbing effects on viewer performance.

At the same time, the use of digital mammograms allows easier measurement of viewer behaviour including the viewing path and the features attracting viewer interest, by means of eye tracking to determine successive locations of viewer gaze during a period of observation time. This is because the digital images need to be viewed on an electronic display screen rather than a light box as used for films, and the eye tracking hardware can be accurately calibrated in this environment. The work reported here makes use of this capability to investigate an important question related to performance of readers, namely the extent to which reader viewing behaviour follows a similar pattern from image to image, and from reader to reader. A method for assessment of reader consistency using these patterns would allow objective comparisons between readers to be made and thereby predict their absolute performance levels and any variations over time. It could also be used on a continuous monitoring basis, to determine whether viewers were affected by performance-reducing effects such as fatigue or distraction.

2 Method

Observations of reader viewing paths for mammograms have been considered by a number of authors previously. Kundel and Nodine (1983) undertook some pioneering studies which linked behavioural habits of readers with sentinel events in the eye tracking sequence (e.g. faster scanning after the first abnormality was detected). They defined a number of overall key parameters for a viewing session which can be extracted from the eye tracking sequence, which have been used by later authors to measure differences in reader behaviour (e.g. “time to first hit”). This work inspired numerous Receiver Operating Characteristic (ROC) studies to be undertaken based on influencing of these different parameters. Krupinski (1996) investigated the difference between intra and inter reader variability using such parameters and concluded that both types of variability were of similar extent. Mello-Thoms (2006) correlated extracted eye tracking parameters with detection performance and observed low variability across multiple readers. None of these authors attempted to use the full eye tracking sequence, which clearly would provide a richer set of values for comparison of different viewers than the subsets of extracted parameters which are typically used.

The work presented here provides a method which can be used to incorporate more information from the eye tracking sequence. The initial basis for this work was the derivation of characterization formulas based on the sequence of gaze positions identified in an eye tracking session, which could be used to distinguish between different observers in a biometric application (Maeder et al 2004). Subsequent work (Maeder and Fookes 2004) applied these formulas to eye tracking data for mammogram readers and reported that inter observer variability tended to be greater than intra. This work in contrast adopts a more independent analytical approach to dealing with the sequence of gaze positions, which is found to lead to a similar conclusion.

The method applied here involves use of a common feature extraction technique employed for pattern analysis based on Principal Component Analysis (PCA). This technique was first utilised in a fully automated face recognition system proposed by Turk & Pentland (1991) to derive a set of face representations which were termed eigenFaces. This technique applies eigen-decomposition to the covariance matrix of a set of M vectorised training sequences of gaze. PCA is used to derive a set of eigenvectors which are ranked based on their eigenvalues λ. The D most relevant eigenvectors are retained to form a sub-space φ. The eigenvalues represent the variance of each eigenvector and so represent the relative importance of each of the eigenvectors with respect to minimising the reconstruction error in a least squares sense. Once the sub-space φ is obtained, a vectorised gaze sequence v can be projected into the space to obtain a feature vector a (a = (vφ - φ0)*φ) where φ0 is the mean gaze vector. This technique is termed “eigenGazes” as each eigenvector is representative of the most variant attributes of the training gaze sequences (similar to eigenFaces as detailed above).

3 Results

The above method was tested on eye tracking data collected from 3 proficient mammogram readers who were presented on 3 separate occasions with the same set of 8 mammograms, each element consisting of paired Medio-Lateral Oblique (MLO) and Cranio-Caudal (CC) projections. 2 of the 8 cases were known positives and the remainder were suspected negatives. The order of the mammograms in each presented set was randomised on each occasion. As only one pair of images could be presented at a time, this test situation is slightly different from the normal reading process, where both MLO and CC images are available for viewing simultaneously. An EyeTech eye tracker (based on infra red computer vision) was used to record the position of the viewer’s point of visual attention on the display screen every 100ms. These points were then scanned to identify successive gaze locations, using a radius of 10 pixels (approximately 2.5 degrees of visual angle) for classifying sequential eye position points as belonging to the same gaze locations. The set of mean positions of all the points for each gaze group provided the gaze location feature values. Figure 1 shows a typical mammogram with the corresponding eye tracker results for a single viewer. For each image, the first 20 gaze locations (including revisits) were extracted and used in the further analysis. Figure 2 shows these results for 5 different sessions with the same viewer.

Principal Component Analysis was then applied to the gaze sequence using the eigenGazes method described.
above. To find the eigenGazes, each gaze sequence was represented as a vector of clustered fixations, \( \Gamma_n \), of length 20. As usual for the construction of a basis set, the mean of the observations is removed and a covariance matrix, \( C \), for the dataset is computed. The eigenGazes then are simply the eigenvectors of \( C \). Using a weighted sum of these eigenGazes, it is possible to reconstruct each gaze location in the dataset. These feature vectors of weights for every gaze sequence were then passed to the classifier to evaluate how well they match within (intra) and between (inter) individual viewers. The results of the dot products of the resulting eigenGazes (on a scale of 0 to 1, with 1 being a perfect match) are shown in Table 1.

<table>
<thead>
<tr>
<th>Class</th>
<th>Viewer(s)</th>
<th>Worst match</th>
<th>Best Match</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intra</td>
<td>1</td>
<td>0.82</td>
<td>0.86</td>
</tr>
<tr>
<td>Intra</td>
<td>2</td>
<td>0.89</td>
<td>0.94</td>
</tr>
<tr>
<td>Intra</td>
<td>3</td>
<td>0.86</td>
<td>0.93</td>
</tr>
<tr>
<td>Inter</td>
<td>1 – 2,3</td>
<td>0.74</td>
<td>0.91</td>
</tr>
<tr>
<td>Inter</td>
<td>2 – 1,3</td>
<td>0.81</td>
<td>0.91</td>
</tr>
<tr>
<td>Inter</td>
<td>3 – 1,2</td>
<td>0.74</td>
<td>0.89</td>
</tr>
</tbody>
</table>

Table 1: PCA based classification of gaze sequences

The intra class results in Table 1 demonstrate that all viewers showed a high degree of “repeatability” in their mammogram viewing behaviour. The match values are high and the ranges between best and worst performance for each viewer are very tight (0.04, 0.05 and 0.07 respectively). Furthermore, the average performance for each viewer is of very small range (0.84 to 0.89). It is interesting to note that the three sets of results are consistent in that a lower worst match implies a lower best match etc. This may be due to other aspects of the viewing process such as speed or concentration, which were not measured during the experiment.

It can also be seen from the above results that the matches for intra classes (in the range 0.82 to 0.94) are slightly higher than those for inter classes (in the range 0.74 to 0.91). At the same time, the average values cover a very small range (0.81 to 0.86) which overlaps strongly with the intra range mentioned above and is of similar size. With data from only 3 viewers it is not reasonable to undertake a confidence analysis of these results, nor to apply conventional methods of statistical comparison (such as kappa), but the considerable overlap suggests that it would not be easy to distinguish an adequate separation between the intra and inter classes.

4 Conclusion

The work reported here provides two outcomes of particular interest for assessment of mammogram reader performance using the eigenGazes approach:

- consistency of gaze sequence performance for a given (intra) reader appears to be sufficiently tight to allow the measurement of deviation from their normal behaviour to be determined;
- consistency of gaze sequence performance between different (inter) readers does not appear to vary substantially, so a common “expected” performance envelope for skilled readers may be able to be determined.

An advantage of the approach adopted here using the eigenGazes PCA assessment method is that a single figure of merit is produced which has intuitive meaning and exhibits monotonic linear characteristics (in a sense similar to a percentage measure).

An immediate extension to this work would be the acquisition of further experimental evidence to increase the validity of the above claims. This work is difficult to perform as access to the time of skilled radiologists for research work is limited. However, national moves towards digital screening mammography in Australia (and elsewhere) and the need to retrain and validate readers in this new environment may offer some opportunity. Some increase in confidence for the results may also be obtainable by comparing them with the behaviour of non-expert viewers, where a greater degree of both intra and inter viewer variability might be anticipated.

The work reported here did not distinguish between gaze sequence behaviour for readers when viewing normal vs abnormal (positive) cases, nor for differing breast morphologies (eg size, texture). Some investigation of the impact on gaze sequence patterns due to attentional attraction in such cases is warranted. This could help improve current computer assisted diagnosis (CAD) software systems for mammogram analysis.

A further matter of interest which could be investigated using the eye tracking data obtained from experiments, is to consider how to compare macro scale aspects of readers’ viewing behaviour (e.g. how much immediate comparison between left and right images they undertake, or whether their overall scanning strategies follow an intended route such as top to bottom). Such aspects will require a more sophisticated clustering approach which extracts information at a higher level of complexity than the individual gaze locations adopted here. This may require the design of specific “signatures” of gaze characteristics which are well suited to the mammogram viewing process rather than to arbitrary image viewing, as those mentioned previously. Nevertheless, the same eigenGazes PCA based method for comparing performance of viewers is a plausible candidate for analysing those situations as well.
5 References


Figure 1: (a) Example of a mammogram (MLO pair) from the test set; (b) one eye tracking sequence for this image.

Figure 2: Examples of sets of the first 20 gaze location clusters identified in 5 independent presentations of the image (m) to the same viewer, following our experimental protocol.