Subjective and computer-based characterisation of mammographic patterns

Kostas Marias[a], Styliani Petroudi[a], Ruth English[b], Rosie Adams[b], Michael Brady[a]

[a] Robotics Research Group, Engineering Science, Oxford University, Ewert House, Ewert Place, Summertown, Oxford, OX2 7BZ, [b] Breast Care Unit, Churchill Hospital, Oxford OX3 7LJ, UK
kostas@robots.ox.ac.uk

Abstract. We investigate the subjective characterisation of breast patterns and the possibility of automatic classification using computed measurements. Several clinical studies used mammogram pattern classification in order to investigate their relation to breast cancer. However, the establishment of an objective ground truth for pattern classification remains a problem that diminishes the potential of such studies since there is significant variability in the interpretation and coding of the patterns. For this reason, we investigate the observer variability in classifying mammographic patterns as well as the variability between film-screen and computer screen reading with respect to pattern classification. We also investigate the role of quantitative measures in characterising the breast parenchyma, discuss their agreement with the observer and argue that a normalised mammographic representation can improve the correspondence between measures and the perceived mammogram patterns.

1. Introduction

In 1976 Wolfe [1] suggested that the mammographic appearance of the breast could be a risk factor for breast cancer. To evaluate this hypothesis he suggested four categories to characterise the breast parenchyma[1]:

N1: The breast consists mostly of fatty tissue with no ducts visible. This represents an essentially normal breast and is considered low risk of cancer.

P1: This represents a fatty breast with predominant ducts in the anterior portion up to 1/4th of the breast. It is also considered low risk.

P2: The breast in involuted with prominent duct pattern of moderate to severe degree, occupying more than 1/4th of the breast volume. The visible duct pattern can occupy the entire breast. It is considered high-risk.

DY: The breast parenchyma is dense, which usually denotes connective tissue hyperplasia[2]. It is considered as the highest risk pattern and can appear homogenous due to the overall increased density. The prominent duct pattern cannot be seen.

[1] The breast parenchyma comprises the lobules, ducts and the interlobular fibrous tissue
[2] The abnormal increase in the number of normal cells in normal arrangement in a tissue

The P2 and DY patterns are the most likely to develop cancer according to Wolfe [1] and since they are denser they can lower mammographic sensitivity or obscure a lesion on mammography. Figure 1, shows an example of Wolfe's patterns.

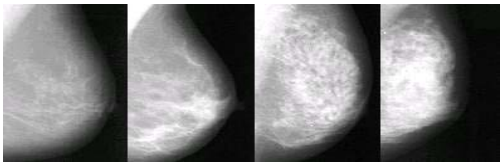


Fig. 1. An example of the Wolfe mammogram patterns (from left to right: N1, P1, P2 and DY)

The American College of Radiologists suggests that breast composition should be reported in all patients using the BI-RADS [2] classification:

I. The breast is almost entirely fat.
II. There are scattered fibroglandular densities.
III. The breast tissue is heterogeneously dense.
IV. The breast tissue is extremely dense.

Mammogram pattern classification can assist clinical studies in defining the role of mammographic appearance as a breast cancer risk. In addition, it can be used to characterise an increase in breast density due to Hormone Replacement Therapy (HRT) that can cause tissue regeneration and even change in breast pattern [3]. However, the classification of mammogram patterns is a subjective task, so it is important to develop an accurate classifier that could automatically report the density class of a given mammogram.

In this paper, we first investigate the observer variability in classifying mammogram patterns (section 2.1). In section 2.2 we discuss the role of quantitative and texture measures for the automatic classification of mammograms, while in section 3 we present some experimental results based on the suggested measures before concluding with a discussion (section 4).

2. Pattern Classification

Before developing method for the automatic classification of the breast parenchyma from the digitised (or digital) mammogram, we consider important to understand the observer variability when using the suggested classification methods (Wolfe, BI-RADS) and the differences when reading directly from the screen vs. film (section 2.1). In section 2.2 we investigate the possibility of computer-based pattern classification.

2.1 Variability in Classification

Two engineers familiar with mammogram patterns and an experienced radiologist classified 132 mammograms using the Wolfe criteria. The radiologist classified the mammograms using both Wolfe and BI-RADS classification. The engineers had

previously formed a consensus for classifying the mammograms blinded to the mammograms used in the experiment. The classification results (from the computer screen reading) of all the observers using the Wolfe criteria are shown in Figure 2.a. It is clear that the engineers' classification is similar while there are significant differences between them and the radiologist especially in the P1 and DY classes.

Figure 2.b shows the radiologist's classifications of the 132 mammograms using both the Wolfe and the BI-RADS criteria. For each case the radiologist classified the mammograms reading them from both the films and the computer screen.

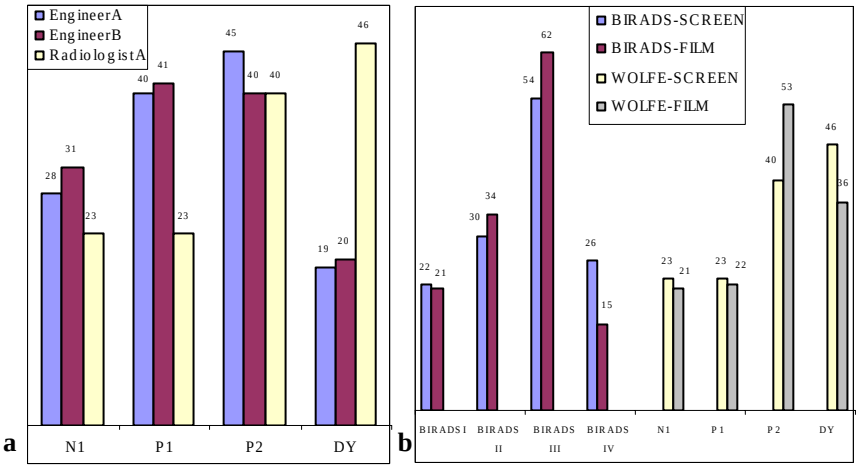


Fig. 2. (a) The classification results of all the observers using the Wolfe criteria, (b) The radiologist classifications using both Wolfe and BI-RADS for both screen and film reading

2.2 Classification Based on Measures

An automatic classifier should depend on robust measures of breast density and texture. Previous efforts to classify mammogram patterns are mostly based on image texture measures (e.g. in [4]). However, differences in imaging conditions lead to complex variations in the mammographic intensity distribution, a fact that diminishes the possible use of texture for mammographic pattern recognition.

For these reasons we use the normalised h_{int} representation of interesting tissue [5] in our experiments. The h_{int} is a quantitative representation of the breast that calculates the height of non-fatty tissue that corresponds to each pixel thus factoring out variations due to differences in image acquisition conditions.

The measures used are:

1. The index volume of "interesting" tissue is defined as:

$$V(h_{\text{int}}) = \frac{\sum_x \sum_y h_{\text{int}}(x, y)}{H \times A} \quad \textbf{(1)}$$

Where A is the segmented mammogram area (excluding the breast edge were the h_{int} is near zero) and H is the height of the compressed breast in each mammogram.

2. Correlation:

$$f_3 = \frac{\sum_i \sum_j ijp(i,j) - \mu_x \mu_y}{\sigma_x \sigma_y} \qquad \mu_{x\,or\,y} = \sum_i i p_{x\,or\,y}(i) \qquad \sigma^2_{x\,or\,y} = \sum_i \left(i - \mu_{x\,or\,y}\right)^2 p_{x\,or\,y}(i) \tag{2}$$

Where $p(i)$ is the probability of grey-level i, and $p(i, j)$ is the 2D probability from the co-occurrence matrix of the image.

3. Results in Automatic Classification

We calculated the measures described in the previous section for the 132 mammograms in our data set. In addition to the V_{int} measure (that is based in the h_{int} representation), we calculate the correlation measure both on the original images and the h_{int} ones. Our principal aim is to understand the significance of various measures for mammogram pattern classification. Since it is difficult (and sometimes unclear to the observer that classifies the patterns) to differentiate P2 from DY (or III and IV in BI-RADS) and P1 from N1 (or I from II), in this experiment we only consider high (P2 and DY or III and IV) vs. low (N1 and P1 or I and II) density classification. Using the V_{int} measure the accuracy in classification was 91% at the V_{int} =21%.

As shown in Figure 3.b, the correlation measure does not show any consistency with the different pattern classes (we use the radiologist's BI-RADS descriptions in this experiment) while in 3.a (h_{int} images) there is a clear separation between the high (III and IV) and low (I and II) density patterns. The accuracy is 88% at f_3=1.7. This clearly indicates the need for normalisation so that texture measures and quantitative measures can become significant and consistent. The result in Figure 3.b is expected since there is no compensation for the variability in imaging conditions and as a result the texture measures are erroneous.

4. Discussion

From the graphs shown in Figure 2, we can conclude that the variability in interpretation and coding is reduced when there is a consensus (as was done by the 2 Engineers). However, as seen in Figure 2.a the perception of the breast patterns can be significantly different and therefore there is a need for a more complete description of these classes as well as the development of standardisation criteria. In addition, as seen in Figure 2.b regardless of the classification method there is a significant difference in film vs. screen classification. Although the classification was performed in a conventional computer screen, the result still addresses an additional issue for the transition to digital mammography and the reading from the screen. Future work in digital mammography may take into consideration the consistency in classifying breast patterns as a mean to evaluate the adequacy of digital systems for mammography.

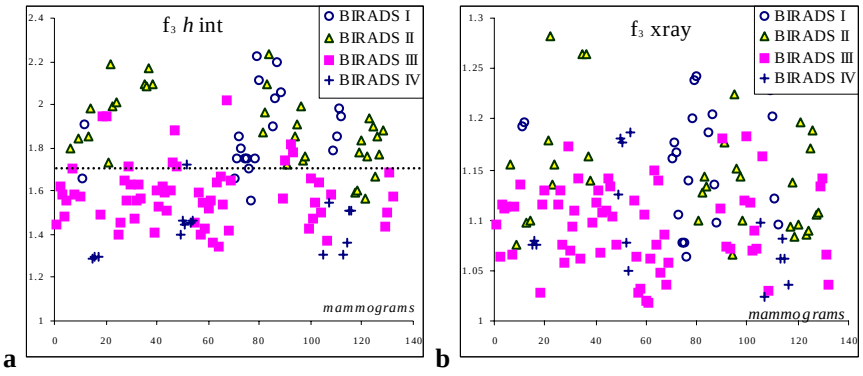


Fig. 3. A simple image texture measure (correlation) calculated in the h_{int} representation of 132 mammograms (a) and in the originals (b). The very good separation between high (III and IV) and low-density (I and II) textures in 3.a makes clear that there is a need for normalisation so that a measure can give significant results.

Finally, the results shown in Figure 3 clearly indicate that the h_{int} normalisation is necessary in order to derive significant measures and develop automatic classifiers of mammogram patterns. So far, our experiments show that a very good discrimination can be achieved between high and low density patterns. However, additional work needs to be done in separating N1 from P1 (or I from II) and P2 from DY (III from IV) and as was discussed in section 2.1 it is necessary to describe patterns more analytically in order to reduce the observer variability and to provide guidelines for the further development of measures that can robustly characterise breast patterns.

Acknowledgements

The authors would like to thank the Cancer Research UK for supporting this work and Mirada Solutions (http://www.mirada-solutions.com/) for the h_{int} generation software.

References

1. Wolfe, J. N.: Breast pattern as an index for developing breast cancer. American Journal of Roentgenology. 126 (1976) 1130-1139
2. ACR.: Illustrated breast imaging reporting and data system (BI-RADS^{TM}). American College of Radiology (1998)
3. Sala, E., Warren, R., McCann. J., Duffy, S., Luben, R., Day, N.: High-risk mammographic parenchymal patterns, hormone replacement therapy and other risk factors: a case-control study. Int J Epidemiol. 29(4) (2000) 629-362
4. Miller, P., Astley, S.: Classification of breast tissue by texture analysis. Image and Vision Computing. 10(5) (1992) 277-282
5. Highnam, R.P., Brady, J.M.: Mammographic image processing. Kluwer Academic Press (1999)