Mammographic image segmentation and risk classification based on mammographic parenchymal patterns and geometric moments*

Wenda He\textsuperscript{a,\dagger}, Erika R.E. Denton\textsuperscript{b}, Kirsten Stafford\textsuperscript{b}, Reyer Zwiggelaar\textsuperscript{a,\ddagger}

\textsuperscript{a} Department of Computer Science, Aberystwyth University, Aberystwyth SY23 3DB, UK
\textsuperscript{b} Department of Radiology, Norfolk & Norwich University Hospital, Norwich NR4 7UY, UK

1. Introduction

In mammographic risk assessment (i.e. the likelihood of a woman developing breast cancer), Tabár et al. proposed a model, based on mixtures of four mammographic building blocks, representing nodular, linear, homogeneous and radiolucent tissue [1]. These four building blocks form the normal breast anatomy. Nodular densities mainly correspond to Terminal Ductal Lobular Units (TDLU); linear densities correspond to either ducts, fibrous tissue or blood vessels, homogeneous-structureless densities correspond to fibrous tissues whose appearance may obscure the underlying normal TDLU and ducts, as well as their alterations due to hyperplastic breast changes; and radiolucent areas are related to adipose fatty tissue, which appears as dark areas in mammographic images [1].

Tabár modelling is strongly influenced by Wolfe's original work, which divided mamograms, using parenchymal patterns, into four risk classes [2]. For Tabár modelling, parenchymal patterns (the relative proportions of tissue belonging to the four building blocks) were used to subdivide mammograms into five risk classes; each of the risk classes is associated with a parenchymal pattern (e.g. a specific distribution of the four Tabár tissue types). Patterns I–V represent low to high mammographic risk [1]. The relative composition of the four building blocks: nodular, linear, homogeneous and radiolucent (respectively) are as follows: Pattern I is composed as \{25\%,15\%,35\%,25\%\}; Pattern II is composed as \{2\%,14\%,2\%,82\%\}; Pattern III is similar in composition to Pattern II, except that the retroareolar prominent ducts are often associated with periductal fibrosis; whilst in Pattern II such an association occurs less frequent. The composition of Pattern IV is [49\%,19\%,15\%,17\%]. Pattern V is composed as \{2\%,2\%,89\%,7\%\} [1]. Fig. 1 shows example mammographic images for Tabár’s five mammographic risk patterns. Between low and high mammographic risk, the probability of developing breast cancer can increase by a factor of 4–37 [1].

* The author would like to acknowledge the financial support for this research through The Research Institute of Visual Computing-Wales (RIVIC).
\dagger Corresponding author.
\ddagger Principal corresponding author.
E-mail addresses: web@aber.ac.uk (W. He), erika.denton@mnuh.nhs.uk (E.R.E. Denton), rrz@aber.ac.uk (R. Zwiggelaar).
To date, mammographic parenchymal pattern classification is based on the subjective appraisal of mammograms, leading to inter and intra observer variability [3,4]. In the case of dense breast parenchyma (e.g. Pattern V), the diagnostic accuracy of both screen- and digital mammography is limited [5]. In order to develop an automated, accurate and repeatable mammographic risk estimation approach based on Tabár risk assessment model, the main challenge is to segment a given mammographic image according to Tabár building blocks, so that the characteristic mixture of these building blocks can be determined, and thus mammographic risk can be estimated [1].

Previous work such as [6,7] have used texture based techniques for mammographic risk classification based on the Breast Imaging Reporting And Data System (BI-RADS) [8]. Also, Refs. [9–11] have focused on risk classification based on Tabár tissue modelling. It should be noted that the BI-RADS scheme describes low to high risk, as the breast tissue is ‘almost entirely fatty’ (e.g. BI-RADS Pattern I) to ‘extremely dense’ (e.g. BI-RADS Pattern IV); whilst Tabár’s scheme is quantitatively defined, based on tissue composition of four mammographic building blocks. He et al. [10] proposed mammographic segmentation using a texton based approach [12] and Tabár tissue modelling. The results indicated that the texton approach can be used to perform mammographic segmentation. However, the evaluation showed over-represented linear structure tissue regions, and under-represented nodular and radiolucent tissue regions in all classes [10]. Similar techniques were used in Ref. [6], in which histogram information (statistical distributions over a texton dictionary) of the whole mammogram was used as a basis for mammographic risk assessment. Oliver et al. [7] employed texture features, after fatty/dense tissue segmentation, directly from mammographic images for automatic classification of breast density. Bayesian combination of two classifiers was able to achieve 86% correct classification for the Mammographic Image Analysis Society (MIAS) database [13] and 77% correct classification for the Digital Database for Screening Mammography (DDSM) [14]. No detailed tissue composition/segmentation results were provided in any of the previous studies [6,7] on risk classification based on BI-RADS scheme.

In the literature, moments have been widely used for shape analysis and object recognition [15]. In this paper, moments are used as statistical measures to represent texture. In the context of texture analysis in mammography, moments have been used for extracting shape related texture features. Application areas include mass and tumour classification [16], microcalcifications classification [17] and mammographic image retrieval [18].

The use of moments is well established in pattern recognition. Based on the mammographic application papers described above, different branches of moments based techniques can be identified. Legendre[19] and Zernike [20] moments are orthogonal basis functions which provide more detailed representation than normalised central moments [15]. Both Legendre and Zernike moments are continuous and have been shown to cope well with image noise.

The use of recurrence relationships to calculate polynomial basis coefficients of discrete moments can lead to numerical error accumulation for the calculation of higher-order moments and larger image sizes, causing significant performance degradation in application [21,22]. The recently proposed Chebyshev moments are discrete polynomial based orthogonal moments which do not involve any kind of approximation, and eliminate discretisation errors whilst preserving all of the theoretical properties, but this is at the cost of computational complexity [23].

Mammographic parenchymal patterns are represented by various breast tissues. The distribution of these tissues contains both shape and periodicity aspects. As an extension of Ref. [24], the proposed mammographic risk classification results are based not only on Tabár but also BI-RADS risk modelling. In addition, this work has incorporated a mammographic segmentation evaluation related to clinical use. The investigation was motivated and supported by the discrete nature of the moments representation, which provides a shape representation, as well as spatial and geometric information (which was absent from the texton based approach [10]). Using geometric moments is appropriate, because the important properties (e.g. shape, kurtosis and skewness) of the Chebyshev moments can be expressed in terms of geometric moments [23], without the computational complexity associated with the Chebyshev moments. In the proposed method the obtained texture feature vectors are expected to contain not only texture primitives like in the texton based approach [12,10], but also geometric information.

2. Materials and methods

The MIAS database [13] was used in the experiments. It contains 322 images, digitised at 50 μm resolution, but only 320 images are usable (file mdb2951l and mdb2961l are excluded for historical reasons). For each of the four Tabár tissue types, an expert mammographic screening radiologist extracted small regions from about 40 randomly selected mammograms. The regions contain typical examples of Tabár building blocks with respect to mammographic parenchymal patterns, resulting in patches containing 199 nodular, 253 linear, 70 homogeneous and 121 radiolucent tissue examples (see the first row in Fig. 1). Of these mammographic patches, 98 were used for detailed annotations (i.e. 21 nodular, 27 linear structure, 28 homogeneous and 22 radiolucent) by the same radiologist, which meant that for each of the patches tissue specific areas were outlined (see the second row in Fig. 2). The collection of mammographic patches were normalised to zero mean and unit variance, and cover the various mammographic risk classes. All the patches were used at the mammographic building block modelling stage, regardless of the associated risk class for the original mammogram. Extending the current work to include alternative databases (e.g. DDSM [14]) would be possible, but would need the manual extraction and detailed tissue outlining of small regions containing typical tissue type examples. This was seen to be outside the scope of this paper.
Ground truth data for each MIAS mammographic image with respect to both Tabár and BIRADS schemes was provided by the same expert mammographic screening radiologist. Both Tabár and BIRADS schemes were used in the mammographic risk classification as a means of determining the robustness of the approach. Note that in both cases (i.e. Tabár and BIRADS), the radiologist performed visual assessment to assign risk classifications, which was based on extensive experience within a national mammographic program and using the relevant documentation [1,8].

The proposed methodology can be broken down into the following distinct steps: (1) feature extraction using mammographic patches, (2) deriving local image properties, (3) feature transformation, (4) mammographic building block based model generation by clustering, and (5) model driven segmentation.

2.1. Local moments calculation

A set of moments, up to order four [18,25], are computed within a small local window centred at each pixel (i,j) in an image (note that image patches are used at the modelling stage and full mammographic images are used at segmentation stage). We used only up to fourth order moments, which are expected to be able to capture sufficient texture geometry information, and at the same time keeps the feature dimensionality low [18,25]. The coordinates of a local window size of $W \times W$ (with $W$ being odd so that the pixel (i,j) is at the centre of the window) are normalised to be in the $[-1,1]$ range. The size of the local windows was determined by Fourier analysis on local patches. The window size was determined by estimating the position of the first side-band in Fourier space [26]. A series of investigations was conducted on parameter settings; to cover the full range of anatomical objects, the chosen sampling regions consisted of four square windows with $W=\{7, 13, 33, 63\}$, where the lower limit indicates the feature extraction of small structures (e.g. small diameter of nodular tissues, thin linear structures), and the upper limit indicates the feature extraction of larger structures. For a pixel $(m,n)$ which falls within the local window $f(m,n)$, the normalised coordinates $(x_m,y_n)$ are defined as $x_m=(m-i)/(W/2)$ and $y_n=(n-j)/(W/2)$. The $(p+q)$th order moments within the window centred at pixel $(i,j)$ are computed by a discrete sum approximation, which uses the normalised coordinates $(x_m,y_n)$, and defined as

$$m_{pq} = \sum_{-W/2}^{W/2} \sum_{-W/2}^{W/2} x_m^p y_n^q f(m,n).$$

(1)

The lower-order moments $(p+q \leq 1)$ have well defined geometric interpretations. The higher-order moments $(p+q \geq 2)$ give more detailed shape characteristics of the polygons [25]. The $m_{00}$ moment can be interpreted as a centre of mass detector, which is aimed at identifying concentrated density (compact high-intensity) at a point. The two first order moments, $m_{10}$ and $m_{01}$, can be interpreted as edge or contrast detectors, which aim to identify image pixels at which the brightness changes sharply or indicate a discontinuity. The second order moment $m_{11}$ can be interpreted as a cross detector, which aims to identify image pixels brightness discontinuities at cross like junctions. Local moments are inherently integral-based features, therefore they reduce the effect of uncorrelated noise. The computation of local moments can be interpreted as a convolution of an image with a set of masks [25], this has been illustrated in Fig. 3. The $x$ and $y$ coordinates of the centre of mass can be computed as $\bar{x} = m_{10}/m_{00}$ and $\bar{y} = m_{01}/m_{00}$ with respect to the normalised coordinates. The central moments are computed for each pixel in the local window, and defined as

$$\mu_{pq} = \sum_{-W/2}^{W/2} \sum_{-W/2}^{W/2} (x_m - \bar{x})^p (y_n - \bar{y})^q f(m,n).$$

(2)

2.2. Deriving local image properties

Local image properties are derived from higher order moments [27,18], and determined as
Variance \( V_{xy} \) (spreading) around the centre of mass with respect to the x-axis and y-axis can be computed from normalised 2nd order moments and defined as

\[
V_{xy} = \frac{\mu_{11}}{\mu_{00}}, \quad V_x = \frac{\mu_{20}}{\mu_{00}} \quad \text{and} \quad V_y = \frac{\mu_{02}}{\mu_{00}},
\]

where \( V_x \) and \( V_y \) are the variances with respect to the x-axis and y-axis, respectively [18].

- Skewness (symmetry) around the centre of mass can be characterised by the normalised 3rd order moments and defined as

\[
S_x = \frac{\mu_{30}}{(\mu_{00}V_x)^{3/2}} \quad \text{and} \quad S_y = \frac{\mu_{03}}{(\mu_{00}V_y)^{3/2}}.
\]

Skewness values of \( S = 0 \), \( S < 0 \) and \( S > 0 \) can be interpreted as a Gaussian (normal) distribution, flatter than a normal distribution, and more peaked than a normal distribution, respectively [18].

- Kurtosis (peakedness) around the centre of mass can be computed from the normalised 4th order moments and defined as

\[
K_x = \frac{\mu_{40}}{(\mu_{00}V_x)^2} - 3 \quad \text{and} \quad K_y = \frac{\mu_{04}}{(\mu_{00}V_y)^2} - 3.
\]

Kurtosis values of \( K = 0 \), \( K < 0 \), \( K > 0 \) and \( K < -1.2 \) can be interpreted as a Gaussian (normal) distribution, flatter than normal distribution, more peaked than normal distribution, and bimodal (multi-modal) distribution, respectively [18].

- The ratio of longest to shortest distance vectors from the centroid of the local window to its boundaries is considered as a measure of elongation of the region and defined as

\[
\text{elongation} = \frac{(\mu_{20} - \mu_{02})^2 + 4\mu_{11}^2}{\mu_{00}}.
\]

- For elongated objects, the orientation \( \theta \) (in degrees) of the major ('long') direction with respect to the x-axis is defined as

\[
\theta = \frac{1}{2} \arctan \frac{2\mu_{11}}{\mu_{20} - \mu_{02}} \times \frac{180}{\pi}.
\]

Thus, for each pixel, 12 attributes were obtained to form a feature vector (i.e., \( \mu_{00}, m_{10}, m_{10}, \bar{x}, \bar{y}, m_{11}, \) elongation, \( \theta, S_x, S_y, K_x \) and \( K_y \)). When taking four scales into account this meant there were 48 features in total representing each pixel.

### 2.3. Feature transformation

Caelli and Oguztoreli [28] introduced a non-linear transformation that mapped moment features to texture features. A hyperbolic tangent function is adopted to map moment feature \( M_k \) with mean \( \bar{M} \) to texture feature \( F_k \) defined as

\[
F_k(i, j) = \frac{1}{L^2} \sum_{(a, b) \in \omega_{ij}} \tanh(\sigma(M_k(a, b) - \bar{M}_k)),
\]

where \( \omega_{ij} \) is an \( L \times L \) (\( L = 55 \)) averaging window centred at location \( (i, j) \), finer textures require a smaller window to detect smaller features, whereas coarser textures require a larger window; the value \( \sigma = 0.01 \) controls the shape of the logistic function; and \( k \in \{1, 2, \ldots, 12\} \). The parameter values were empirically determined in order to achieve optimal results, and small variations did not significantly affect the results. Example moment images before and after transformation are shown in Fig. 4. Before the non-linear transformation, the extracted feature images can be interpreted visually as

![Example moment images](image-url)
line detection with different orientations and scales; the enhanced feature images show increased differentiation on tissue specific areas. This is expected to lead to more discriminative feature vectors.

2.4. Mammographic model generation and segmentation

Once the transformed feature vectors have been obtained, Tabár mammographic building blocks are modelled using K-means clustering (K=10). A total of 40 cluster centres are generated to represent nodular, linear, homogeneous and radiolucent tissue (10 for each of the four tissue types). The number of cluster centres (K value) was determined empirically based on the assumption that each mammographic building block contains at least 4–6 texture primitives according to their visual appearance covering different orientations and scales. A smaller number of cluster centres can lead to less discriminative models; conversely, too many cluster centres can lead to many similar models and redundancy. Visual assessment of the segmentation of mammographic patches using the manual annotations (e.g. outlined tissue specific areas) provided by the expert radiologist, was used to heuristically decide the number of cluster centres. It should be noted that there is no rigorous way to determine exactly how many texture primitives should be used for complex texture patterns as seen in mammographic images [10]. However, the segmentation results of patches (see Fig. 2) did indicate a strong correlation with the provided tissue outlines. In addition, the set of features does describe all mammographic anatomical aspects covering lines and regions at various scales.

The computer vision literature has shown that a low intensity background (e.g. low or zero grey level values) can lead to meaningless moments [25]. A homogeneous high intensity background can also lead to meaningless moments. When performing texture analysis in digital mammography, such aspects can cause misclassification between high density homogeneous tissue and low density radiolucent tissue. Intensity of homogeneous and radiolucent areas may vary between different mammographic risk patterns. In addition, image normalisation can alter the intensity distribution. To tackle the problem of misclassification due to meaningless moments between high intensity tissues (e.g. homogeneous and nodular) and low intensity tissues (e.g. radiolucent), a threshold post-processing was incorporated into the classifier to reduce classification errors. In particular, this involves reclassifying tissue that is initially classified as radiolucent, but has a very high intensity, to homogeneous or nodular tissue. The threshold values were determined based on prior knowledge of the intensity distribution and variation of mammographic images across the whole MIAS database, and defined as the mean value of homogenous and nodular tissues from the collection of annotated mammographic patches. It should be noted that in practice, mammographic images may be obtained through different image acquisition processes. Therefore, for an alternative database, the threshold values may need to be re-evaluated using a set of training images obtained based on a specific image acquisition process.

At the segmentation stage, for each pixel of a mammographic image, the same procedure was applied to obtain transformed feature vectors. For each pixel, the resultant feature vector was compared to the distribution of all the mammographic building block models. A distance weighted K-Nearest-Neighbour (KNN) classification (K=9) was used; to assign a Tabár mammographic building block class to each pixel, which weighted the contribution of each of the K nearest neighbours with the Mahalanobis distance [29] to the query point, giving greater weight to closer neighbours to reduce misclassification. The K value was determined empirically, but variation in these parameters indicated robustness.

2.5. Evaluation

A four-step evaluation was employed: (1) perform segmentation on annotated mammographic patches to verify that the proposed methodology is adequate for producing realistic segmentation on tissue specific areas, (2) perform a quantitative and qualitative evaluation on mammographic image segmentation, using the MIAS database, (3) conduct mammographic risk classification based on the tissue composition from the segmented images, and (4) clinical evaluation of the segmentations to assess the realistic and practical use of the proposed methodology in a clinical environment.

Customised software was developed for clinical evaluation purposes, which allowed a radiologist to grade a given mammographic segmentation from five available options (i.e. unacceptable, poor, acceptable, good and very good). In addition, the grades can be combined and referred to as principal grades; the first two grades are considered as clinically unacceptable (CUA), the remaining three are considered as clinically acceptable (CA). An expert consultant radiologist performed the evaluation. The statistical results are presented as percentages with respect to the five grades and principal grades, and associated with both Tabár and BIRADS risk classifications, as well as the corresponding low and high risk categories. In addition, expert feedback was collected and incorporated in the investigation, including segmentation accuracy on tissue specific areas, and misclassified breast tissues associated with the segmentation methodology issues. This information is taken into account to describe the relationship between the risk classification achieved and clinical practice satisfaction, so as to determine the effectiveness of the segmentation methodology and the possibility of developing it for clinical utilisation in screening mammography or computer aided diagnosis systems.

3. Experimental results

This section is arranged in line with the four evaluation stages as described in Section 2.5.

3.1. Mammographic patch and image segmentation

The segmentation on patches was visually assessed as a preliminary validation, where the segmented patches were directly compared with the annotated data (manual outlines of tissue specific regions within the patches). Example segmentations are shown in Fig. 2, which indicate realistic segmentation on tissue specific areas. Comparing the manual annotations with the automatic segmentations, the later provides more details, have a tendency to over segment linear structures, and exclude (under segments) small areas from nodular and homogeneous regions. Segmentation on mammographic images showed improved segmentation accuracy when compared with textron based approaches [10,30], in terms of the structure of breast parenchymal pattern as well as the relative proportion of the mixture of four building blocks. Fig. 5 shows an example segmented mammographic image. The results indicate the difference in the relative area of the various tissue types, and their distribution across the mammographic image.

Table 1 shows the average relative proportion of the four building blocks, and their standard deviation with respect to Tabár tissue modelling, based on all the segmented images from the MIAS database. These results show an improved distribution in all four risk patterns, when compared with the textron based approach [30]. However, Table 1 shows that on average radiolucent tissue is under represented in Pattern II/III. In Pattern II/III and V the relative proportion of homogeneous tissues differs from the expected value;
both cases have relatively large standard deviations, indicating large intra class variation in the segmentation.

3.2. Risk classification

The resultant tissue composition from each segmented mammographic image was used as a feature vector for the risk classification. The ground truth was provided by an expert radiologist (see Section 2).

With respect to Tabár based risk classification, the feature vectors were compared with the five Tabár models (risk patterns), as described in Ref. [1] (see Section 1) and using a Euclidean distance. Table 2 shows the classification accuracy when discriminating between Tabár categories. The results show good overall classification, except for Tabár Pattern I (see discussion below).

To determine the robustness of the proposed approach, the risk classification was repeated based on Tabár and BIRADS risk categories [8], but using a nearest neighbour (e.g. segmented mammographic image) classifier with a Euclidean metric, and a leave-one-image-out evaluation methodology. Table 3 shows classification for discriminating between the four BIRADS risk categories using the ground truth data. The results based on Tabár risk categories (detailed confusion matrix not included) also indicated a total-four class accuracy of 70% and a high/low class accuracy of 79%, but the correct classification for Tabár Pattern I was only 55%.

The results show a clear correlation between the Tabár and BIRADS results, except for Tabár Pattern I. It should be noted that recent research showing a comparison between Wolfe, Boyd, BIRADS and Tabár based mammographic risk assessment [4] indicated strong correlations between Wolfe, Boyd and BIRADS whilst Tabár based assessment only shows similar correlation when Tabár Pattern I is excluded [4].

3.3. Clinical evaluation

Table 4 shows the segmentation assessment results, which for Tabár I, at 42% is especially low; however, the clinical acceptable rate increases as the risk level gets higher which is particularly encouraging, with the Tabár V class reaching over 96% clinical acceptable results. For BIRADS risk classes I-III, the clinical acceptable rate for segmentation was about 55%; and 92% clinical acceptable rate was achieved for BIRADS risk class IV. It should be noted that the number of samples for Tabár V or BIRADS IV are small. These aspects are also clear for the low/high risk results, which

Table 1
Showing the average relative proportion (as percentages) of the four building blocks, and their standard deviation with respect to Tabár tissue modelling.

<table>
<thead>
<tr>
<th></th>
<th>Nodular</th>
<th>Linear</th>
<th>Homogeneous</th>
<th>Radiolucent</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>26 ± 18</td>
<td>14 ± 5</td>
<td>25 ± 17</td>
<td>32 ± 21</td>
</tr>
<tr>
<td>II/III</td>
<td>13 ± 10</td>
<td>8 ± 4</td>
<td>21 ± 20</td>
<td>56 ± 23</td>
</tr>
<tr>
<td>IV</td>
<td>38 ± 19</td>
<td>12 ± 5</td>
<td>25 ± 15</td>
<td>23 ± 14</td>
</tr>
<tr>
<td>V</td>
<td>12 ± 18</td>
<td>3 ± 4</td>
<td>67 ± 27</td>
<td>16 ± 13</td>
</tr>
</tbody>
</table>

Table 2
Classification confusion matrix based on Tabár risk categories. Total accuracy was 53%, \( \kappa = 0.36 \) (fair agreement). Total accuracy for low and high categories was 71%, \( \kappa = 0.40 \) (fair agreement).

<table>
<thead>
<tr>
<th>Tabár pattern</th>
<th>Automatic</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>I</td>
<td>II/III</td>
</tr>
<tr>
<td>Truth</td>
<td>33</td>
<td>36</td>
</tr>
<tr>
<td>II/III</td>
<td>14</td>
<td>66</td>
</tr>
<tr>
<td>IV</td>
<td>22</td>
<td>6</td>
</tr>
<tr>
<td>V</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 3
Classification confusion matrix based on BIRADS risk categories. Total accuracy was 70%, \( \kappa = 0.59 \) (moderate agreement). Total accuracy for low and high categories was 79%, \( \kappa = 0.58 \) (moderate agreement).

<table>
<thead>
<tr>
<th>BIRADS Pattern</th>
<th>Automatic</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>I</td>
<td>II</td>
</tr>
<tr>
<td>Truth</td>
<td>65</td>
<td>10</td>
</tr>
<tr>
<td>II</td>
<td>0</td>
<td>71</td>
</tr>
<tr>
<td>III</td>
<td>0</td>
<td>18</td>
</tr>
<tr>
<td>IV</td>
<td>0</td>
<td>5</td>
</tr>
</tbody>
</table>
Table 4
Clinical grading of satisfaction regarding the mammographic segmentation. ‘U’, ‘P’, ‘A’, ‘G’ and ‘VG’ denote unacceptable, poor, acceptable, good and very good, respectively.

<table>
<thead>
<tr>
<th>#imgs</th>
<th>U</th>
<th>P</th>
<th>A</th>
<th>G</th>
<th>VG</th>
<th>CA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tabár</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>119</td>
<td>31%</td>
<td>27%</td>
<td>16%</td>
<td>18%</td>
<td>8%</td>
</tr>
<tr>
<td>II/III</td>
<td>93</td>
<td>23%</td>
<td>19%</td>
<td>22%</td>
<td>22%</td>
<td>14%</td>
</tr>
<tr>
<td>IV</td>
<td>80</td>
<td>14%</td>
<td>17%</td>
<td>24%</td>
<td>25%</td>
<td>20%</td>
</tr>
<tr>
<td>V</td>
<td>28</td>
<td>4%</td>
<td>0%</td>
<td>14%</td>
<td>21%</td>
<td>61%</td>
</tr>
<tr>
<td>Low</td>
<td>212</td>
<td>27%</td>
<td>24%</td>
<td>19%</td>
<td>19%</td>
<td>11%</td>
</tr>
<tr>
<td>High</td>
<td>108</td>
<td>11%</td>
<td>13%</td>
<td>21%</td>
<td>24%</td>
<td>31%</td>
</tr>
<tr>
<td>BIRADS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>87</td>
<td>19%</td>
<td>24%</td>
<td>29%</td>
<td>25%</td>
<td>3%</td>
</tr>
<tr>
<td>II</td>
<td>103</td>
<td>28%</td>
<td>16%</td>
<td>20%</td>
<td>17%</td>
<td>19%</td>
</tr>
<tr>
<td>III</td>
<td>93</td>
<td>24%</td>
<td>24%</td>
<td>16%</td>
<td>21%</td>
<td>15%</td>
</tr>
<tr>
<td>IV</td>
<td>37</td>
<td>2%</td>
<td>6%</td>
<td>21%</td>
<td>21%</td>
<td>50%</td>
</tr>
<tr>
<td>Low</td>
<td>190</td>
<td>24%</td>
<td>19%</td>
<td>23%</td>
<td>22%</td>
<td>12%</td>
</tr>
<tr>
<td>High</td>
<td>130</td>
<td>20%</td>
<td>20%</td>
<td>17%</td>
<td>21%</td>
<td>22%</td>
</tr>
</tbody>
</table>

show a relative better clinical acceptable result for high risk categories, where the high risk results based on Tabár risk modelling are promising as over 76% are considered clinically acceptable results.

Further details of the segmentation assessment results can be found in Table 5, where the classification results from Tables 2–4 are correlated. Table 5 shows the number of cases which were correctly classified (i.e. all the cases on the diagonal of the confusion matrices in Table 2) in the mammographic risk assessment experiments against the clinical evaluation of the segmentation results. Results for mammographic risk assessment that were misclassified by a single, two and three classes (e.g. ±1 to II would be I and III) are also included (note that misclassification by three classes only happens for a few cases). The results in Table 5 show a strong correlation between the level of correctness of the mammographic risk classification and clinical evaluation of the segmentation results: i.e. correct mammographic risk classification is associated with the good and very good segmentation results, whilst misclassification by two and three classes is mainly associated with the poor and unacceptable segmentation results.

4. Discussion and conclusions

We have presented a method to perform mammographic segmentation based on geometric moments, which is shown to produce realistic segmentation results with respect to tissue specific areas. The results also show good segmentation accuracy and overall good classification results, except for Tabár Pattern I. This latter aspect can be linked to previous published results [4], which are consistent with our classification results, as Tabár Pattern I shows weak correlation between BIRADS and Tabár risk models.

Incorrect segmentation can lead to misclassification. This is particularly sensitive when performing risk classification based on Tabár tissue modelling, because the tissue composition from a segmented image is used as the feature vector. It is interesting to note that improved classification results were achieved based on the BIRADS risk classification scheme.

A direct comparison with existing mammographic risk assessment methods is not always possible (mainly due to differences in databases) [6,11]. The size of the database, resolution of the images, and the variation in the images (e.g. across the BIRADS density classes) can influence the classification results. A direct comparison with Oliver et al. [7] is probably the closest option. In Ref. [7] the MIAS database was used, 82% correct classification was obtained based on ‘expert C’ [7]; the same ground truth data was used in this study. However, it should be noted that the proposed mammographic segmentation is specifically designed with respect to Tabár mammographic parenchymal patterns, and has fundamental differences to density based segmentation as described in Ref. [7]. In this study, BIRADS risk classification was obtained to facilitate the classification performance evaluation when using different risk models (Tabár vs. BIRADS), based on mammographic parenchymal patterns rather than density information alone.

At the feature extraction stage, moments were calculated up to fourth order which is expected to be able to capture sufficient geometry information [18]. The current implementation uses a local square window for feature extraction. Despite this, the sampling region may not be sufficient to capture some extremely larger tissue structures, square sampling windows are likely to include a mixture of tissues (e.g. linear structure patches used in the training can also contain radiolucent tissue), which is not ideal for modelling a specific texture. A model selection stage could be incorporated in the methodology to tackle this issue [31].

Some of the misclassification may be caused by low contrast regions (e.g. radiolucent tissue) leading to meaningless moments. Also, high intensity homogeneous density regions (e.g. homogeneous tissue) can lead to a similar problem. This indicates that the currently used texture features may miss out some discriminative intensity information. A threshold based post-processing step is used to reduce the false positives.

The decreasing segmentation accuracies in the individual low risk classes indicates that the precision of the segmentation are not all up to clinical standard; from a clinical point of view, this shows that a mathematically correct risk classification (possibly due to a robust classifier), does not necessarily reflect the correctness of the associated segmentation. It is encouraging to see the accuracies increased in high risk classes (i.e. Tabár V and BIRADS IV), but also interesting to notice small classification variation between BIRADS risk classes I, II and III. This may indicate that the presented methodology is able to discriminate mammographic parenchymal patterns well based on Tabár risk modelling, but less robust in extracting density information (possible due to meaningless moments) which is the key feature of BIRADS based risk models.

The segmentation assessment results seem to indicate an apparent difference between the way in which radiologists perceive mammographic risk and how the segmentation performs. The automatic segmentation of radiolucent tissue seems less robust. However, in associating the correct classification of mammograms with the assessment of the segmentation results, Table 5 indicates a strong correlation between these. This may also indicate that the tissue modelling associated with the segmentation process does not cover the various tissue classes appropriately and there might be a strong non-linear component in the expert assessment (e.g. the areas of linear and radiolucent tissue play a less important role

Table 5
Clinical grading of satisfaction regarding the mammographic segmentation associated with the level of automatic mammographic risk classification. ‘LH’ denotes low and high risk.

<table>
<thead>
<tr>
<th>#imgs</th>
<th>U</th>
<th>P</th>
<th>A</th>
<th>G</th>
<th>VG</th>
<th>CA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tabár</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correct</td>
<td>170</td>
<td>11%</td>
<td>21%</td>
<td>22%</td>
<td>24%</td>
<td>22%</td>
</tr>
<tr>
<td>±1</td>
<td>68</td>
<td>16%</td>
<td>13%</td>
<td>22%</td>
<td>29%</td>
<td>20%</td>
</tr>
<tr>
<td>±2</td>
<td>72</td>
<td>50%</td>
<td>21%</td>
<td>14%</td>
<td>8%</td>
<td>7%</td>
</tr>
<tr>
<td>±3</td>
<td>10</td>
<td>40%</td>
<td>40%</td>
<td>20%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>LH correct</td>
<td>227</td>
<td>11%</td>
<td>20%</td>
<td>22%</td>
<td>26%</td>
<td>21%</td>
</tr>
<tr>
<td>LH incorrect</td>
<td>93</td>
<td>48%</td>
<td>20%</td>
<td>13%</td>
<td>10%</td>
<td>9%</td>
</tr>
<tr>
<td>BIRADS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correct</td>
<td>224</td>
<td>16%</td>
<td>16%</td>
<td>23%</td>
<td>25%</td>
<td>20%</td>
</tr>
<tr>
<td>±1</td>
<td>68</td>
<td>24%</td>
<td>32%</td>
<td>13%</td>
<td>15%</td>
<td>16%</td>
</tr>
<tr>
<td>±2</td>
<td>17</td>
<td>58%</td>
<td>24%</td>
<td>12%</td>
<td>0%</td>
<td>6%</td>
</tr>
<tr>
<td>±3</td>
<td>11</td>
<td>82%</td>
<td>9%</td>
<td>9%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>LH correct</td>
<td>253</td>
<td>15%</td>
<td>19%</td>
<td>22%</td>
<td>25%</td>
<td>19%</td>
</tr>
<tr>
<td>LH incorrect</td>
<td>67</td>
<td>48%</td>
<td>25%</td>
<td>10%</td>
<td>7%</td>
<td>10%</td>
</tr>
</tbody>
</table>
when the areas of dense tissues are significant), which is currently not part of the tissue modelling approach.

The major concerns from the radiologist who participated in the clinical evaluation include: (1) pectoral muscle and upper abdominal fat are often misclassified as nodular or homogeneous fat within nodular tissues; (2) radiolucent areas are often misclassified as dense tissue (see Fig. 6); (3) misclassification of tissue between nodular and homogeneous classes may be caused by (apparently) under exposed films (see Fig. 7). Clinical feedback indicates some segmentation methodology issues, such as the lack of discriminative modelling between nodular and homogeneous tissues, as well as between nodular and radiolucent tissues. In addition, it seems necessary to have an approach designed specifically for dealing with intensity variation as seen on digitised mammograms; the normalisation method used may be too general and in the current approach can have a direct effect on the model generation. Intra and inter observer variation is noticeable [4]; however, it should be noted that the issue of experts’ subjective assignment of risk classification can be dealt with using multiple readers, and the majority risk classifications can be considered
as ‘final’, but full evaluation of these aspects is seen as future work.

Currently, breast density is considered as a stronger indicator in early breast cancer detection than parenchymal patterns, and is favoured in screening mammography, as the sensitivity of mammography is significantly reduced by increased breast density because of the masking of some tumours due to dense fibro glandular tissue [32]. However, work to fully understand the density-risk association is still ongoing. At the same time, parenchymal pattern segmentation (four classes) has a clear advantage over dense tissue segmentation (two classes), as clinical evidence has shown that early cancers can be seen during retrospective or temporal analysis on changes in parenchymal patterns of mammograms taken prior to cancer being detected; such changes may provide more information than having percent density alone [33]. Automatic mammographic segmentation is also expected to be useful as a means of aiding radiologists’ diagnosis and treatment planning prior to biopsies.

In terms of computational efficiency, the lab experiments indicate that each segmentation takes a matter of minutes to process on a personal computer (i.e. Window OS, Intel Pentium 4 and 4 GB RAM). Clinical implementation may require separate investigations on the software and hardware architectures, in order to optimise the process time.

So in summary, in this paper, the proposed geometric moments based methodology has shown a way to extract texture features in mammographic images, which are used for mammographic image segmentation. The presented results show realistic segmentation on tissue specific areas, improved estimation of the relative proportion of the four building blocks when compared to a texton based approach, leading to promising mammographic risk classification results. In addition, a clinical evaluation was performed to confirm the effectiveness of the methodology in a clinical setting, and show strong positive correlations with the automatic risk classification. This paper provides links between mammographic segmentation, geometric moments, density information, parenchymal pattern information, and mammographic risk assessment based on Tabár or BI-RADS schemes. Classification accuracies of 71% and 79% were achieved in the corresponding low and high risk categories for Tabár and BI-RADS schemes, respectively. The presented clinical evaluation is based on the entire MIAS database, which is publicly available. The reported comprehensive results can be considered as benchmark results for direct comparison with respect to a follow up clinical evaluation to ensure realistic and practical use in a clinical environment.

References