Adapting Breast Density Classification from Digitized to Full-Field Digital Mammograms

Meritxell Tortajada[1], Arnau Oliver[1], Robert Martí[1], Mariona Vilagran[2], Sergi Ganau[2], Lidia Tortajada[2], Melcior Sentís[2], and Jordi Freixenet[1]

[1] Computer Vision and Robotics,
University of Girona, Girona, Spain
txell@eia.udg.edu
[2] Department of Breast and Gynecological Radiology,
UDIAT-Diagnostic Center, Parc Taulí Corporation, Sabadell, Spain

Abstract. Mammographic density is strongly associated with breast cancer, being considered one of the most important risk indicators for the development of this type of disease. Likewise, the sensitivity of automatic breast lesion detection systems is significantly dependent on breast tissue characteristics. Therefore, the measurement of density is definitely useful for detecting breast cancer. The aim of this work is to adapt our previously developed automatic breast tissue density classification methodology for digitized mammograms to full-field digital mammograms (FFDM), as well as to evaluate the possible improvements and the classification results. After breast area extraction and peripheral enhancement, the method segments the breast area into fatty and dense tissue, then morphological and texture features from each class are extracted and finally FFDM are classified according to a standard qualitative criteria. Results show a strong correlation ($\kappa = 0.88$) between automatic and expert assessments and a better classification correction percentage (CCP = 92%) compared to our earlier work.

Keywords: Breast density classification, full-field digital mammography, feature extraction and selection, peripheral enhancement.

1 Introduction

Mammographic density represents the amount of radiodense tissue within the breast and it is one of the strongest risk factors for breast cancer. Most of the studies about the relationship between breast density and breast cancer report that women with high dense breast have greater risk of breast cancer than those with low dense breast [2]. Besides risk of developing breast cancer, density is also related to the difficulty of detecting breast cancer [13]. The latest studies show that even though breast density does not affect the sensitivity of microcalcification detection Computer Aided Detection (CAD) systems, it significantly affects mass detection, so the sensitivity of CAD systems for mass detection decreases

A.D.A. Maidment, P.R. Bakic, and S. Gavenonis (Eds.): IWDM 2012, LNCS 7361, pp. 561–568, 2012.

in dense mammograms [15]. Therefore, breast density assessment is regarded as an important tool to help radiologists and CAD systems to detect breast cancer.

Mammographic density can be measured both quantitatively and qualitatively. Quantitative studies use the estimation of the percentage of breast density (dense area divided by total area), the absolute dense area or the breast density volume [8]. Whereas for qualitative assessment, the Wolfe categories, the Tabár grade or the Breast Imaging Reporting and Data System (BIRADS) score [12] can be used. However, BIRADS classification is becoming a standard on the evaluation of mammographic density where four patterns are used: (I) the breast is almost entirely fat ($< 25\%$ glandular), (II) there are scattered fibroglandular densities ($25 - 50\%$ glandular), (III) the breast tissue is heterogeneously dense ($51 - 75\%$ glandular) and (IV) the breast tissue is extremely dense ($> 75\%$ glandular).

We have previously presented a breast tissue density classification methodology for digitized mammograms [14]. The main purposes of this work are to extend our method for digitized mammograms to FFDM, to assess the benefits of the updates, and finally to classify digital mammograms according to the BIRADS score.

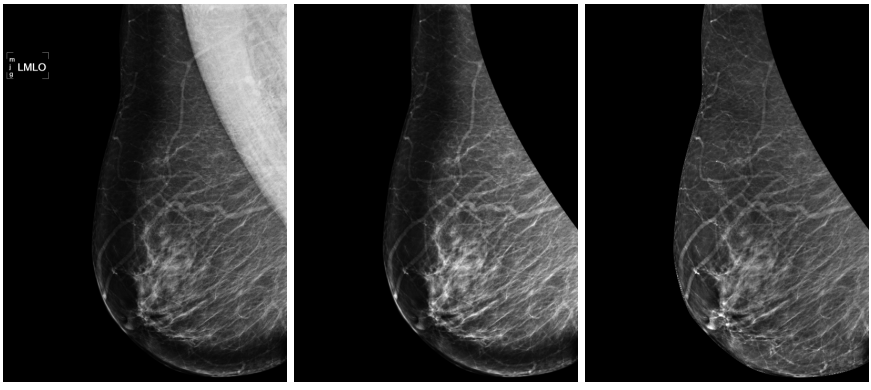
2 Original Methodology

The classification method is based on our previously developed algorithm for breast tissue density classification [14]. The original method consisted in: (1) preprocessing, (2) segmentation in fatty and dense tissue, (3) feature extraction from both classes, and (4) classification according to BIRADS categories.

(1) **Preprocessing:** During the preprocessing step, the breast skin-line and the pectoral muscle are detected using the approach of Kwok et al. [11]. Mammograms are divided in breast area, background and pectoral muscle, and only the breast area is kept.
(2) **Segmentation:** Gray-level information in combination with the fuzzy C-means (FCM) clustering approach is used to group the pixels of the breast area into fatty and dense tissue classes.
(3) **Feature extraction:** Once the breast area is divided into two classes, a set of morphological and texture features for fatty and dense tissue are extracted. As morphological features, the relative area and the first four moments of the histogram are calculated and as texture features, the ones derived from co-occurrence matrices.
(4) **Classification:** BIRADS classification of mammograms is performed using three classifiers: K-Nearest Neighbor (KNN), the C4.5 decision tree and a Bayesian classifier based on the combination of KNN and C4.5.

3 Updated Methodology

Although the general idea of the methodology is preserved, some changes have been done to adapt the method to FFDM. The main differences are: (1) a peripheral enhancement is applied during the preprocessing stage and (2) additional feature selection techniques and classifiers are tested during the classification stage.



(a) (b) (c)

Fig. 1. Example of the preprocessing process: (a) original image, (b) breast area segmentation and (c) peripheral enhancement

3.1 Peripheral Enhancement

The first results of the FCM segmentation were not accurate enough due to the presence of an overexposed area in the majority of mammograms (see Fig. 1(a)-(b) and Fig. 2(a)). This is a known issue that happens during mammographic acquisition because of breast thickness changes. During mammographic acquisitions, breast is compressed with a tilting compression paddle, so the breast thickness can be non uniform, being lower in the periphery and overexposing this area. We decided to compensate the thickness variations in the periphery of the breast by a peripheral enhancement method that is similar to the work of Karssemeijer et al. [10] but using a multiplicative model. After extracting the breast boundary and the pectoral muscle, the overexposed area is determined by Otsu's thresholding and a correction factor is applied over each pixel of the detected region. To calculate the correction factor, firstly a distance map is generated using the distance from each point (x) in a mammogram (M) to the breast skin line. From the furthest peripheral pixel to the closest, each pixel value $M(x)$ at distance i is divided by the mean value of its neighborhood at distance i ($N_i(x)$) and multiplied by the mean value of its neighborhood at distance $i+1$ ($N_{i+1}(x)$), where $N_i(x) = \{t \in M \ \textit{at distance}\ i : \textit{distance}(t,x) \leq k\}$, being k an experimentally set parameter (100 for our case). An example of the overall process can be seen in Fig.1.

3.2 Feature Selection and Classification

Due to the large number of features, a feature-selection step is included selecting the most effective subset of features. Various feature selection techniques are evaluated (using WEKA [18] data mining software) such as Principal Component Analysis [9], Gain Ratio attribute evaluation [17] or Support Vector Machine (SVM) [6]. The classification of mammograms according to BIRADS categories

is also performed with WEKA. We used more recent classifiers like Random Forest [3] or SVM [4] and some combinations of classifiers as AdaBoost [5] or a binary tree of SVM. The binary tree consists in firstly, classification of digital mammograms in low or high breast density category and then low dense cases are classified in BIRADS I or II and independently, high dense cases in BIRADS III or IV. The reason was to convert our multiclass classification problem into multiple binary classification problems as SVM is originally a binary classifier [1].

4 Results

The method was applied to the whole set of 236 FFDM acquired with a Selenia FFDM system that form our local digital database. This database is composed of left or right Medio-Lateral Oblique mammograms from 236 healthy women.

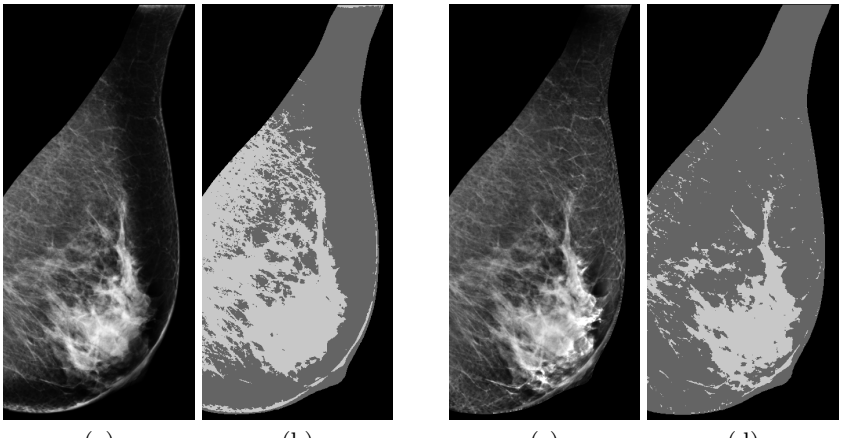


Fig. 2. Example of the segmentation process: (a) original breast area, (b) FCM without previous peripheral enhancement, (c) breast area after peripheral enhancement and (c) FCM with previous peripheral enhancement.

4.1 Preprocessing and Segmentation

To determine not only the quality of the preprocessed images but also the segmentation results, visual assessment was performed by one observer with more than 10 years of experience in mammographic images. To evaluate the enhancement process, the observer labeled the images as correctly enhanced or not and a total of 83% of the images were considered to be improved with the peripheral enhancement. An example of the enhancement results can be seen in Fig. 1(c). To assess the segmentation improvements, the observer evaluated the differences in the segmentation results when images were enhanced or not. Around 92% of the segmentations obtained after image enhancement were considered similar or

better than the ones obtained before enhancement. Specifically, the 45% were regarded as strictly better, therefore results show that there is a clear improvement in the segmentation results when images are previously peripheral enhanced (see Fig. 2).

4.2 Classification

Experts Classification. Four expert mammographic readers classified all the images using BIRADS (current readers are different from the ones that participated in the study [14]). The ground truth was determined by majority vote. In case of tie, the median value was considered as the consensus opinion (like in [14]).

Table 1(A)-(D) shows the confusion matrices for the classification of FFDM for the four readers and the consensus opinion in year 2011. Like in our previous work [14], the results show an evident interobserver variability, illustrating the difficulty of the breast tissue density classification task. In low dense breasts categories {BIRADS I & II}, expert B tends to classify in BIRADS I (17 mammograms were classified as BIRADS I being BIRADS II) whereas experts C and D tend to classify in BIRADS II (31 and 34 mammograms respectively were classified as BIRADS II being BIRADS I). Note also that expert B repeats this underestimation assignment when classifying in BIRADS II (8 mammograms were classified as BIRADS II being BIRADS III) and expert C repeats the overestimation appointment when classifying in BIRADS III (11 mammograms were classified as BIRADS III being BIRADS II). In high dense categories {BIRADS III & IV}, expert D differs from the rest considering a few BIRADS III mammograms (18/46) and a lot of BIRADS IV (27 mammograms were classified as BIRADS IV being BIRADS III). When considering the individual BIRADS classes, the correct classification percentage (CCP) values for expert A are really high (99%, 98%, 85%, 100%, respectively). The results of the other experts are less homogeneous and lower, except for expert C in BIRADS III with CCP = 91%. Using the Cohen's kappa coefficient (κ) values, the agreement of expert A with the consensus opinion belongs to the *almost perfect* category ($\kappa = 0.94$) whereas the agreement of experts B, C and D with the consensus opinion belong to the *substantial* category ($\kappa = 0.78, 0.70, 0.61$ respectively).

Furthermore, a few years ago, one of the experts classified the same database according to BIRADS. Table 1(D)-(E) shows the confusion matrices for the classification of FFDM for one reader and the consensus opinion, in two different periods of time. Results reveal intraobserver variability in BIRADS II and III classification. In the past the reader classified 88 mammograms as BIRADS II whereas now the number increases to 120. On the other hand, 52 mammograms were considered BIRADS III opposite to the current 20. Examining each class, there are no significant variations in CCP values for BIRADS I (before: 58%, after: 60%), BIRADS III (before: 37%, after: 39%), and BIRADS IV (before: 94%, after: 100%), opposite to the CCP values for BIRADS II (before: 59%, after:97%).

Table 1. (A)-(D) Confusion matrices for four expert radiologists and their consensus opinion and (E) confusion matrix for one expert radiologist and the consensus opinion in 2005.

Consensus	Expert A (Year 2011) $\kappa = 0.94$ $CCP = 96\%$				Expert B (Year 2011) $\kappa = 0.78$ $CCP = 86\%$				Expert C (Year 2011) $\kappa = 0.70$ $CCP = 79\%$				Expert D (Year 2011) $\kappa = 0.61$ $CCP = 73\%$				Expert D (Year 2005) $\kappa = 0.41$ $CCP = 57\%$			
	I	II	III	IV	I	II	III	IV	I	II	III	IV	I	II	III	IV	I	II	III	IV
I	84	1	0	0	85	0	0	0	54	31	0	0	51	34	0	0	49	36	0	0
II	1	86	1	0	17	67	4	0	0	77	11	0	0	85	2	1	0	52	35	1
III	0	0	39	7	0	8	36	2	0	2	42	2	0	1	18	27	0	0	17	29
IV	0	0	0	17	0	0	4	13	0	0	3	14	0	0	0	17	1	0	0	16
	(A)				(B)				(C)				(D)				(E)			

Automatic Classification. To evaluate our algorithm, we used a leave-one-out methodology, i.e., each digital mammogram is analyzed by a classifier trained using the mammograms of all other women in the database. Table 2(C) shows the best confusion matrix after analyzing different feature selection and classification methods. Specifically the confusion matrix is obtained using SVM feature selection followed by binary tree of SVM classification and this combination achieved a κ of 0.88 and a CCP of 92% (216/236). These values are higher than the values of experts B, C and D although they are lower than the ones of the expert A. When considering each BIRADS classes, the CCP for BIRADS I is 93% (79/85), for BIRADS II is 89% (78/88), for BIRADS III is 93% (43/46) and for BIRADS IV is 94% (16/17). Note that BIRADS III reaches the highest CCP value in comparison with the ones reached by the experts (A: 85%, B: 78%, C: 91%, D: 39%).

Table 2. Confusion matrices for MIAS, DDSM and digital databases classification and their respectively consensus opinion: (A) Bayesian combination of KNN and C4.5 classifiers in MIAS, (B) Bayesian combination of KNN and C4.5 classifiers in DDSM and (C) SVM Selection (SVS) + Binary Tree of SVM classification (BTSVC) in digital database.

Consensus	Bayesian MIAS $\kappa = 0.81$ $CCP = 86\%$				Bayesian DDSM $\kappa = 0.67$ $CCP = 77\%$				SVS + BTSVC $\kappa = 0.88$ $CCP = 92\%$			
	I	II	III	IV	I	II	III	IV	I	II	III	IV
I	79	1	3	4	58	25	23	0	79	6	0	0
II	3	86	6	8	15	295	26	0	5	78	5	0
III	0	2	85	8	12	46	196	1	0	2	43	1
IV	0	6	4	27	5	18	18	93	0	0	1	16
	(A)				(B)				(C)			

Table 2(A)-(B) also shows the best confusion matrices of our work in [14]. In this case the used classifier was a Bayesian combination of KNN and C4.5 classifiers and the classification method was tested using two public databases: the Mammographic Image Analysis (MIAS) database [16] and the Digital Database for Screening Mammography (DDSM) [7] which were obtained from scanned or digitized film images. Although a direct comparison with our previous results is difficult because the datasets used are different, in principle, the confusion matrix of the digital database (Table 2(C)) seems to be better than the others because there are less non-zeros off-diagonal elements. When comparing κ and CCP values in digital and digitized databases ($\kappa = 0.81$, CCP = 86% (277/322) for MIAS and $\kappa = 0.67$, CCP = 77% (642/831) for DDSM), they are slightly better in the digital case. Examining the individual BIRADS classes, the CCP for MIAS data set were 91%, 84%, 89% and 73% (respectively) and for DDSM were 55%, 88%, 77% and 69% (respectively). All these values are also somewhat better in the digital case (93%, 89%, 93% and 94%), although the highest difference is in BIRADS IV. Using the two-class classification (low vs high density), the CCP for low case is 97% for digital, 89% for MIAS and 89% for DDSM, whereas for high case is 97% for digital, 94% for MIAS and 79% for DDSM, so in both cases, the percentage is higher for digital database. These results make explicit the improvement reached with the updated method.

5 Conclusions

We have provided a breast density classification method that can be applied to both digitized and digital mammograms. For our digital database we obtained a κ of 0.88 and a CCP of 92% that represents a better agreement in 3 out of 4 radiologists. Results are also better than our previous work using MIAS and DDSM, which indicates that the included changes improve the overall method. In the future, we plan to work in two directions: (1) although the segmentation results are qualitatively better when including peripheral enhancement, other segmentation algorithms will be investigated; and (2) regions of interest will be described using other texture features.

Acknowledgment. M. Tortajada holds a UdG grant BRGR10-04. This work was partially supported by the Spanish Science and Innovation grant TIN2011-23704.

References

1. Allwein, E.L., Schapire, R.E., Singer, Y.: Reducing multiclass to binary: A unifying approach for margin classifiers. J. Mach. Learn. Res. 1, 113–141 (2000)
2. Boyd, N.F., Lockwood, G.A., Byng, J.W., Tritchler, D.L., Yaffe, M.J.: Mammographic densities and breast cancer risk. Cancer Epidem. Biomarkers Prev. 7(12), 1133–1144 (1998)
3. Breiman, L.: Random forests. Mach. Learn. 45(1), 5–32 (2001)

4. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. 20(3), 273–297 (1995)
5. Freund, Y., Schapire, R.E.: Experiments with a new boosting algorithm. In: Int. Conf. Mach. Learn. / Conf. Uncert. Art. Intell. / Conf. Learn. Theory, pp. 148–156 (1996)
6. Guyon, I., Weston, J., Barnhill, S., Vapnik, V.: Gene selection for cancer classification using support vector machines. Mach. Learn. 46, 389–422 (2002)
7. Heath, M., Bowyer, K., Kopans, D., Moore, R., Kegelmeyer, P.J.: The Digital Database for Screening Mammography. In: Int. Work. Dig. Mammography, pp. 212–218 (2000)
8. Highnam, R., Brady, S.M., Yaffe, M.J., Karssemeijer, N., Harvey, J.: Robust Breast Composition Measurement - Volpara^{TM}. In: Martí, J., Oliver, A., Freixenet, J., Martí, R. (eds.) IWDM 2010. LNCS, vol. 6136, pp. 342–349. Springer, Heidelberg (2010)
9. Jolliffe, I.T.: Principal Component Analysis, 2nd edn. Springer (2002)
10. Karssemeijer, N., te Brake, G.M.: Combining single view features and asymmetry for detection of mass lesions. In: Int. Work. Dig. Mammography, pp. 95–102 (1998)
11. Kwok, S.M., Chandrasekhar, R., Attikiouzel, Y., Rickard, M.T.: Automatic pectoral muscle segmentation on mediolateral oblique view mammograms. IEEE Trans. Med. Imag. 23(9), 1129–1140 (2004)
12. American College of Radiology. Illustrated Breast Imaging Reporting and Data System BIRADS, 3rd edn. American College of Radiology (1998)
13. Oliver, A., Freixenet, J., Martí, J., Pérez, E., Pont, J., Denton, E.R.E., Zwiggelaar, R.: A review of automatic mass detection and segmentation in mammographic images. Med. Image Anal. 14(2), 87–110 (2010)
14. Oliver, A., Freixenet, J., Martí, R., Pont, J., Pérez, E., Denton, E.R.E., Zwiggelaar, R.: A novel breast tissue density classification methodology. IEEE Trans. Inform. Technol. Biomed. 12(1), 55–65 (2008)
15. Romero, C., Varela, C., Cuena, R., Almenar, A., Pinto, J.M., Botella, M.: Impact of mammographic breast density on computer-assisted detection (CAD) in a breast imaging department. Radiología 53(5), 456–461 (2011)
16. Suckling, J., Parker, J., Dance, D.R., Astley, S.M., Hutt, I., Boggis, C.R.M., Ricketts, I., Stamatakis, E., Cerneaz, N., Kok, S.L., Taylor, P., Betal, D., Savage, J.: The Mammographic Image Analysis Society digital mammogram database. In: Int. Work. Dig. Mammography, pp. 211–221 (1994)
17. Witten, I.H., Frank, E.: Data Mining Pactical Machine Learning Tools and Technique, 3rd edn. Morgan Kaufmann (2005)
18. Witten, I.H., Frank, E., Hall, M.A.: Data Mining: Practical machine learning tools and techniques, 3rd edn. Morgan Kaufmann (2011)