ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence
and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

NMF-Density: NMF-Based Breast Density Classifier

Lahouari Ghouti*and Abdullah H. Owaidh

King Fahd University of Petroleum and Minerals - Department of Information
and Computer Science. KFUPM Box 1128. Dhahran 31261 - Saudi Arabia.

Abstract. The amount of tissue available in the breast, commonly characterized by the breast density, is highly correlated with breast cancer. In fact, dense breasts have higher risk of developing breast cancer. On the other hand, breast density influences the mammographic interpretation since it decreases the sensitivity of breast cancer detection. This sensitivity decrease is due to the fact that both cancerous regions and tissue appear as white areas in breast mammograms. This paper introduces new features to improve the classification of breast density in digital mammograms according to the commonly used radiological lexicon (BI-RADS). These features are extracted from non-negative matrix factorization (NMF) of mammograms and classified using machine learning classifiers. Using ground truth mammographic data, the classification performance of the proposed features is assessed. Simulation results show that the latter significantly outperforms existing density features based on principal component analysis (PCA) by achieving higher classification accuracy.

1 Introduction

Digital mammography represents an efficient means for breast cancer detection. Computer-aided diagnosis (CAD) of breast cancer has emerged as a support tool for medical radiologists in the early detection of breast cancer. On the other hand, recent medical studies have revealed a strong correlation between the mammographic density and the risk of breast cancer [6]. However, breast density can negatively influence the decision of radiologists since the detection of cancer tumors can be obstructed by the tissue density [6]. This paper aims at providing a mechanism for systematic classification of breast densities according to the BI-RADS lexicon. It is hoped that this automated mechanism will facilitate the work of radiologists in the mammographic evaluation for breast cancer detection. The proposed density classification scheme introduces new features based on the non-negative matrix factorization (NMF) technique. In the literature, several approaches are proposed for the automated classification of breast density. These approaches can be classified into: 1) matrix factorization; 2) global histogram; and 3) texture analysis methods. Matrix factorization techniques factorizes the mammogram images into a product of several *factor* images according to specific constrains. Consequently, the mammographic images, known for their high dimensionality, undergo a drastic dimensionality reduction where only dominant features are kept. Oliver et al.[6] proposed a two-class

*Research supported by King Abdul-Aziz City for Science and Technology through NSTP Project Project No. 11-MED1612-04.

breast density classification. Image segmentation is used as a pre-processing step. Then, features are extracted using principle component analysis (PCA) and linear discriminant analysis (LDA) techniques to classify the mammogram images into *fatty* and *dense* types. Features extracted using 2D-PCA are proposed by De Olivera et al. [3] to build a two class (fatty and dense) content-based image retrieval (CBIR) system. A support vector machine (SVM) with Gaussian kernels classify image features represented by the first four principle components (PC). Reported results indicate that 2D-PCA outperforms the standard PCA in terms of classification accuracy. Using the same features, proposed in [3], Thomas et al. [4] consider 4 density classes according to the BI-RADS lexicon using a similar classifier. De Oliveira et al. [1] propose a CBIR system, called *MammoSVD*, where image features are extracted using the singular value decomposition (SVD) algorithm. It is noteworthy that *MammoSVD* system is a binary classifier (fatty and dense tissue) based on an SVM learning machine. The SVD-based features provide a good characterization of the mammographic texture. *MammoSVD* system achieves 90% classification accuracy. In [2], a 4-class model, called *MammoSVx* is proposed where features are represented using the largest 25 singular values of the SVD decomposition of the mammogram images. Using an SVM learning model with polynomial kernel against a mammographic database containing 10000 images, a classification accuracy of 82.14% is achieved by *MammoSVx.* This paper is organized as follows. Section 2 provides a detailed description of the NMF technique. The proposed scheme for the classification of the breast density is discussed in Section 3 where the features extracted from mammographic images using the NMF factorization are also introduced. Section 4 gives a summary of the performance analysis carried out to assess the accuracy of the proposed breast density classification scheme. The paper concludes with Section 5 where conclusions and directions for future work are given.

2 Non-Negative Matrix Factorization (NMF)

NMF is an unsupervised learning approach that leads to parts-based image representations. Such representations are generated using additive combinations of the original images [1]. Also, the *non-negativity constraint* imposed on the factorization allows for more realistic extracted image factors [5]. Given a non-negative input image, $\mathbf{A} \in \mathbb{R}_{\neq}^{m\times n}$, the NMF yields the following factorization:

$$\mathbf{A} \approx \mathbf{W} \times \mathbf{V} \tag{1}$$

where the rows of $\mathbf{W} \in \mathbb{R}_{\neq}^{m\times r}$ and the columns of $\mathbf{V} \in \mathbb{R}_{\neq}^{r\times n}$ represent the NMF basis and their encoding coefficients, respectively. Image approximation is achieved using ranks satisfying: $(m+n)\ r < m \times n$. Keeping in mind that the NMF does not allow negative entries in $\mathbf{W}$ and $\mathbf{V}$, it has found several applications including face recognition and gene expression. Fig. 1 reveals the power of NMF factorization in terms of the locality of the features extracted. It is clear that the NMF bases are well-localized unlike the PCA ones, which gives

[1] Other factorization methods such as PCA and independent component analysis (ICA) yield subtractive combinations leading to *holistic* image representations.

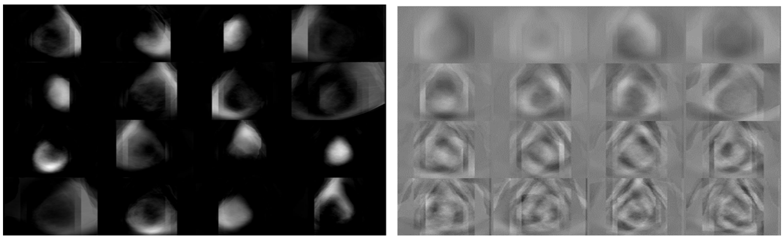

Fig. 1: NMF versus PCA decomposition. First 16 NMF bases (left). First 16 PCA bases (right).

NMF-based features more discriminating capability. The factorization, given by Eq. 1, defines the following optimization problem: Given a non-negative image, $\mathbf{A} \in \mathbb{R}_{\neq}^{m\times n}$, find non-negative approximations, $\mathbf{W} \in \mathbb{R}_{\neq}^{m\times k}$ and $\mathbf{V} \in \mathbb{R}_{\neq}^{k\times n}$, such that $k < \min(m, n)$. Then, this non-convex constrained optimization is defined as follows:

$$f\left(\mathbf{W},\mathbf{V}\right) = ||\mathbf{A} - \mathbf{W}\mathbf{V}|\,|_2^F = \sum_{ij}\left(\mathbf{A}_{ij} - \left(\mathbf{W}\mathbf{V}\right)_{ij}\right)^2 \tag{2}$$

The Frobenius norm, $||\cdot|\,|_2^F$, is used to measure the approximation error. Other common objective functions include the well-known *Kullback-Leibler* divergence objective function:

$$D\left(\mathbf{A}||\mathbf{W}\mathbf{V}\right) = \sum_{ij}\left(\mathbf{A}_{ij}\log\frac{\mathbf{A}_{ij}}{\left(\mathbf{W}\mathbf{V}\right)_{ij}} - \mathbf{A}_{ij} + \left(\mathbf{W}\mathbf{V}\right)_{ij}\right) \tag{3}$$

Eq. 3 can be solved using different algorithms including multiplicative updates, gradient descent and alternating least squares [5]. The multiplicative updates, proposed by Lee and Seung [5], for solving Eq. 2, are given by:

$$\begin{aligned}\mathbf{W}_{ij} &\longleftarrow \frac{\left(\mathbf{A}\mathbf{V}^T\right)_{ij}}{\left(\mathbf{W}\mathbf{V}\mathbf{V}^T\right)_{ij}}\mathbf{W}_{ij}\\ \mathbf{V}_{ij} &\longleftarrow \frac{\left(\mathbf{W}^T\mathbf{A}\right)_{ij}}{\left(\mathbf{W}^T\mathbf{W}\mathbf{V}^T\right)_{ij}}\mathbf{V}_{ij}\end{aligned} \tag{4}$$

3 Proposed NMF-Based Breast Density Classification System

The proposed NMF-Based breast density classification system is shown in Fig. 2. This system consists of 3 main building blocks:

1. **Preprocessing and segmentation:** The preprocessing step is crucial for successful and error-free mammographic interpretation. This step includes noise removal and contrast enhancement. The pectoral muscle, visible in MLO views, is segmented apart enabling the extraction of the image region of interest (ROI). In the case of the experimental database used in this paper, the extracted ROIs contain 300 × 300 pixels. A sample

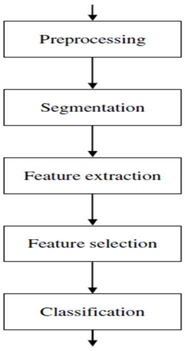

Fig. 2: Proposed Block-Based Breast Density Classification Scheme.

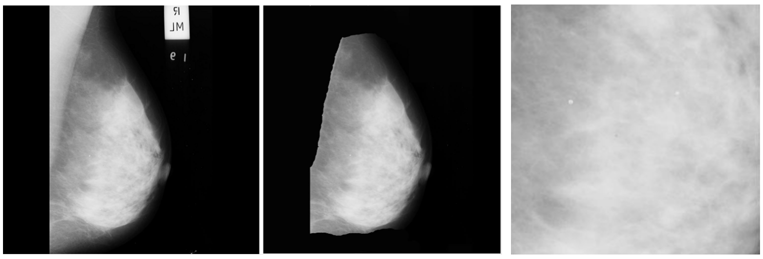

Fig. 3: Raw mammogram image (left). Preprocessed and segmented sample (middle). 300 × 300 ROI area (right).

mammographic image and its preprocessed sample are shown in Fig. 3. It is obvious that preprocessing and segmentation have significantly improved the visual quality of the image prior to inspection by radiologists.

2. **Feature extraction and selection:** Mammogram images, having the same density annotation, are grouped into a large mammogram image which is decomposed using the NMF factorization given by Eq. 1. Features are then extracted by retaining only the first few factors. The NMF factorization efficiency is illustrated in Fig. 1 where only 16 NMF factors are retained.
3. **Machine learning-based classification:** Given their universal classification capabilities, support vector machines (SVM) are proposed to classify the breast density classes (binary or multi class). As such, a CBIR system can be developed based on the breast density categorization. In general, the SVM classifier finds the linear decision boundary (or hyperplane) that successfully separates data pertaining to two given classes. Moreover, this hyperplane maximizes the separating distance between the two classes. higher classification performance is achieved by greater separating distance.

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7. Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

Features	Full Mammogram	Patches (ROIs)
PCA (first 5 components)	50.62	55.62
PCA (first 10 components)	50.31	57.81
NMF (first 5 factors)	72.43	77.84
NMF (first 10 factors)	75.34	**83.19**

Table 1: Mean accuracy of PCA- and NMF-based density classification schemes using full mammogram and ROI images.

4 Simulation Results and Discussion

To assess the capability of the proposed NMF-based breast density classification, the mammographic image analysis society (MIAS) database is used. This public database, freely available, contains 322 mammographic images (50 micron pixel edge) taken in medio-lateral oblique (MLO) view [7]. All images, reduced to 1024 × 1024 pixels, were annotated by experienced radiologists and classified into three distinct density categories: 1) fatty (106 images); 2) fatty-glandular (104 images); and 3) dense-glandular (112 images). Training and testing experiments were conducted using 70% and 30% of the full mammogram and ROI images, respectively [2]. Breast densities are also classified according to the BI-RADS lexicon. This lexicon describes the breast density, along with the lesion feature and classification. Table 1 summarizes the classification results of the conducted experiments. Second and third columns of this table show respectively the mean accuracy obtained from the testing phase using 30% of the data. Simulation results reported for the PCA-based features are reproduced from [7]. As it can be clearly observed, NMF yielded higher classification accuracy thanks to its parts-based factorization. Also, the use of ROIs allowed higher classification accuracy than using full mammography. In fact, the use of ROIs with NMF yielded the best classification accuracy since the extracted local features were not biased towards the background regions which are usually dominant in mammographic images. Finally, it is cautionary to mention that increasing the number of retained factors does not always yield higher accuracies, This is reminiscent of the *over-fitting problem* often encountered in artificial neural networks.

5 Conclusions

In this paper, we proposed a new classification scheme for breast density using non-negative matrix factorization (NMF) and support vector machines. Unlike other factorization techniques, NMF enables the extraction of local mammographic features thanks to the non-negativity constraint imposed. Also, the iterative solution of the NMF makes it an excellent candidate for mammogram images which are usually characterized by high resolution and depth. The performance of the NMF-based breast density classification scheme is assessed using

[2] Pre-processed full mammogram and ROI images are available at: https://github.com/welber/cad_mammography/tree/master/MIAS_300_300.

a standard, annotated and publicly-available mammographic database. As reported in this paper, the proposed classification scheme does not only achieve higher classification accuracy but can properly handle the invariance in mammogram images due to the breast density as well. Finally, the proposed density classification is based on *local (or parts)-* based features which yield sparse structures when the sparsity constraint is imposed on the NMF factors. Future work includes a detailed investigation of the classification capability of the proposed algorithm versus industry-acclaimed density estimation tools such as the *VolparaDensity* (previously known as *Cumulus*) tool. Also, extending the proposed scheme to classify all BI-RADS density types will be considered.

Acknowledgment

The author would like to acknowledge the support provided by King Abdulaziz City for Science and Technology (KACST) through the Science & Technology Unit at King Fahd University of Petroleum & Minerals (KFUPM) for funding this work through project No. 11-MED1612-04 as part of the National Science, Technology and Innovation Plan.

References

[1] J. E. E. de Oliveira, G. Camara-Chavez, A. de Araujo, and T. M. Deserno. Mammosvd: A content-based image retrieval system using a reference database of mammographies. In *22nd IEEE International Symposium on Computer-Based Medical Systems*, pages 1–4, 2009.

[2] J. E. E. de Oliveira, G. Camara-Chavez, A. de Araujo, and T. M. Deserno. content-based image retrieval applied to bi-rads tissue classification in screening mammography. *World Journal of Radiology*, 3(1):24–31, 2011.

[3] J. E. E. de Oliveira and A. de Araujo. Mammosyslesion: A content-based image retrieval system for mammographies. In *17th International Conference on Systems, Signals and Image Processing (IWSSIP 2010)*, pages 408–411, 2010.

[4] T. M. Deserno, M. Soiron, J. E. E. de Oliveira, and A. de Araujo. Towards computer-aided diagnostics of screening mammography using content-based image retrieval. In *24th Conference on Graphics, Patterns and Images (Sibgrapi 2011)*, pages 1754–1760, 2011.

[5] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[6] A. Oliver, X. Lado, E. Perez, J. Pont, J. Denton, E. Freixenet, and J. Marti. Statistical approach for breast density segmentation. *Journal of Digital Imaging*, 23(5):55–65, 2009.

[7] W. R. Silva and D. Menotti. Classification of mammograms by the breast composition. In *International Conference on Image Processing, Computer Vision, and Pattern Recognition (ICPV 2012)*, pages 1–6, July 2012.