Wavelet-based Feature Analysis for Classification of Breast Masses from Normal Dense Tissue

Filippos Sakellaropoulos, Spyros Skiadopoulos, Anna Karahaliou, George Panayiotakis, and Lena Costaridou
Department of Medical Physics, School of Medicine,
University of Patras, 265 00 Patras, Greece
costarid@upatras.gr

Abstract. Automated detection of masses on mammograms is challenged by the presence of dense breast parenchyma. The aim of this study was to investigate the feasibility of using wavelet-based feature analysis for differentiating masses, of varying sizes, from normal dense tissue on mammograms. The dataset analyzed consists of 166 regions of interest (ROIs) containing spiculated masses (60), circumscribed masses (40) and normal dense tissue (66). A set of ten multiscale features, based on intensity, texture and edge variations, were extracted from the ROIs subimages provided by the overcomplete wavelet transform. Logistic regression analysis was employed to determine the optimal multiscale features for differentiating masses from normal dense tissue. The classification accuracy in differentiating circumscribed masses from normal dense tissue is comparable with the corresponding accuracy in differentiating spiculated masses from normal dense tissue, achieving areas under the ROC curve 0.895 and 0.875, respectively.

1 Introduction

Breast cancer is the most prevalent cancer among women [1]. While screen/film mammography is currently the primary imaging technique for early detection and diagnosis of breast cancer, its high diagnostic performance is challenged by occult disease signs (masses and/or microcalcifications) due to the masking effect of dense breast parenchyma, often both characterized by quite similar radiographic densities [2,3]. While microcalcification clusters are indicative of early malignant processes, masses are the most important signs for detection of invasive breast cancer, with their extent being a very important prognostic factor. Masses can be described as more or less compact areas that appear brighter (radiopaque) than the parenchymal

Please use the following format when citing this chapter:
Sakellaropoulos, Filippos, Skiadopoulos, Spyros, Karahaliou, Anna, Panayiotakis, George, Costaridou, Lena, 2006, in IFIP International Federation for Information Processing, Volume 204, Artificial Intelligence Applications and Innovations, eds. Maglogiannis, I., Karpouzis, K., Bramer, M., (Boston: Springer), pp. 722–729

tissue. The major morphological categories of masses are spiculated and circumscribed [4].

Various features in combination with classification methods have been proposed for automated mass detection. Kegelmeyer et al. [5] have introduced edge orientation features based on local edge orientation histogram analysis as well as Laws' texture energy measures to identify spiculated mass containing areas. Chan et al. [6,7] proposed multiresolution texture analysis extracted from spatial Gray Level Dependence Matrices (GLDM) for differentiation of masses from normal tissue. Later on, Liu et al. [8] extended mass edge orientation analysis with a multiresolution scheme for the detection of spiculated masses. Linear phase non-separable 2D wavelet transform (WT) was used to extract features at each resolution on a pixel basis. Petrick et al [9] and Kobatake et al. [10] have utilized a combination of boundary (morphological) and multiresolution texture features (GLDM analysis) to identify and segment the extent of masses, respectively. Another approach in differentiating mass containing areas from normal tissue refers to area patterns constructed using principal component, independent component and factor analyses [11,12].

The performance of the proposed mass detection methods is characterized by high sensitivity (84-96%) and is challenged by the high number of false positive detections per image (1.0-4.4), especially in case of dense tissue [13,14].

The aim of this study is to investigate discriminant features for mass detection in a demanding mass dataset. To capture significant information from intensity, texture and edges of masses of various sizes and to differentiate them from normal dense tissue, histogram, texture and orientation-based features were extracted from the coefficients of an overcomplete wavelet transform. Stepwise forward logistic regression analysis was employed to determine the most discriminating subset of features in differentiating: (a) spiculated masses from normal dense tissue, (b) circumscribed masses from normal dense tissue and (c) both types of masses from normal dense tissue. The performance of the logistic regression models is evaluated by means of Receiver Operating Characteristic (ROC) analysis.

2 Materials and Methods

2.1 Case Sample

Mammographic images corresponding to extremely dense or heterogeneously dense (density 3 and 4, according to BIRADS lexicon) originating from the Digital Database for Screening Mammography (DDSM) of the University of South Florida [15] were selected. Images were digitized with Lumisys or Howtek scanner, at 12 bits pixel depth with spatial resolution of 50 μm and 43.5 μm, respectively. Regions of interest (ROIs) were selected with an image visualization tool developed in our department [16]. The sample consists of 166 ROIs, 60 ROIs containing spiculated masses, 40 ROIs containing circumscribed masses and 66 ROIs of normal dense

tissue. The mean size (longest dimension) was 19 mm (range: 7-49 mm) and 12 mm (range: 6-31 mm) for spiculated and circumscribed masses, respectively. Histogram of mass subtlety (from 1=subtle to 5=obvious), according to DDSM database, is provided in Figure 1.

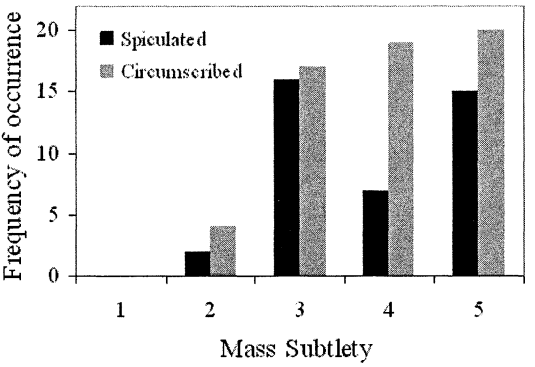

Fig. 1. Histogram of subtlety for spiculated and circumscribed masses of the sample.

2.2 Wavelet Transform

A fast, biorthogonal, Redundant Discrete Wavelet Transform (RDWT), originally used to derive multiscale edges from signals, is utilized in this work. It is based on a family of wavelet functions $\psi(x)$ with compact support, which are derivatives of corresponding Gaussian-like spline functions $\theta(x)$. The algorithm is implemented using a filter bank algorithm, called "algorithm a trous" [17,18], which does not involve subsampling. The discrete wavelet transform is a uniform sampling of the wavelet transform series, discretized over the scale parameter s at dyadic scales 2^j (wavelet transform series). [17,18]. The RDWT is calculated up to a coarse dyadic scale J. Therefore, the original image is decomposed into a multiresolution hierarchy of subband images, consisting of a coarse approximation image $S_{2^J} f(m,n)$ and a set of wavelet images $\left(W^1_{2^j} f(m,n), W^2_{2^j} f(m,n)\right)_{1 \le j \le J}$, which provide the details that are available in $S_1 f$ but have disappeared in $S_{2^J} f$. All subband images have the same number of pixels as the original, thus the representation is highly redundant. The RDWT computes the multiscale gradient vector. Coefficient subband images are proportional to the sampled horizontal and vertical components of the multiscale gradient vector, and thus they are related to local contrast [19]. The magnitude-orientation representation of the gradient vector, in the discrete case, is given by:

$$M_{2^j}(m,n)=\sqrt{\left|W_{2^j}^1(m,n)\right|^2+\left|W_{2^j}^2(m,n)\right|^2}\ ,\ A_{2^j}(m,n)=\arctan\left(\frac{W_{2^j}^2(m,n)}{W_{2^j}^1(m,n)}\right) \quad (1)$$

2.3 Feature Extraction

The aim was to capture significant information from intensity, texture and edges of masses for their differentiation from normal dense tissue. Information from intensity, texture and edges from each ROI, containing spiculated or circumscribed mass, was extracted by means of histogram, texture and orientation features, respectively.

2.3.1 Histogram-based Features

Significant information from intensity variations was extracted by computing features based on histogram of low-frequency band (approximation image) for each ROI at three resolution scales (three to five). These features depend only on individual pixel values and not on the interaction or co-occurrence of neighboring pixel values. In this study, four features corresponding to the four moments of each ROI subimage histogram were calculated: Mean value, Standard Deviation, Skewness and Kurtosis.

2.3.2 Texture-based Features

Laws' texture energy features have been computed by first applying small convolution kernels to ROI subimages, and then performing a nonlinear windowing operation. In this study, the most discriminant set of convolution kernels [5] were used: L5*E5, E5*S5, L5*S5 and R5*R5, for texture discrimination between masses and normal dense tissue.

2.3.3 Orientation-based Features

To capture significant information from mass edges, two multiscale gradient-orientation features were extracted from each ROI:

- Standard Deviation of Gradient-Orientation [5,8]:

$$\sigma_{\text{hist}}(\text{i},\text{j})=\sqrt{\frac{1}{255}\sum_{n=0}^{255}\left(\text{hist}_{\text{ij}}(\text{n})-\overline{\text{hist}_{ij}}\right)^2} \quad (2)$$

where $hist_{ij}$ is the histogram of gradient orientations θ, calculated from high-frequency wavelet coefficients (detailed orientation image) for three scales (three to five), within $N(i,j)$ using 256 orientations (bins). Therefore, $hist_{ij}(n)$ is the number of pixels in $N(i,j)$ that have gradient orientations $\theta\in(-\pi/2+\text{n}\,\pi/256,-\pi/2+(\text{n}+1)\pi/256)$ where $n=0,1,2,\ldots,255$. $\overline{\text{hist}_{\text{ij}}}=\frac{1}{256}\sum_{n=0}^{255}\text{hist}_{\text{ij}}(\text{n})$ is the average bin height of $hist_{ij}$.

- Standard Deviation of Folded Gradient-Orientation [8]:

$$\sigma_\beta(i,j)=\sqrt{\frac{1}{K-1}\sum_{(m,n)\in N(i,j)}\left(\beta(m,n)-\overline{\beta(i,j)}\right)^2} \quad (3)$$

where the folded gradient orientation $\beta(i,j)$ is defined as:

$$\beta(i,j)=\begin{cases}\theta(i,j)+\pi & \text{if } \overline{\theta_+}(i,j)-\theta(i,j)>\dfrac{\pi}{2} \text{ and } KP\geq KN\\ \theta(i,j)-\pi & \text{if } \theta(i,j)-\overline{\theta_-}(i,j)>\dfrac{\pi}{2} \text{ and } KP<KN\\ \theta(i,j) & \text{otherwise}\end{cases} \tag{4}$$

where

$$\overline{\theta_+}(i,j)=\frac{1}{KP}\sum_{\theta(m,n)\geq 0,(m,n)\in N(i,j)}\theta(m,n),\ \overline{\theta_-}(i,j)=\frac{1}{KN}\sum_{\theta(m,n)\leq 0,(m,n)\in N(i,j)}\theta(m,n) \tag{5}$$

are the mean values of positive and negative gradient orientations within $N(i,j)$, respectively. KP and KN are the number of positive and negative gradient orientations within $N(i,j)$, respectively.

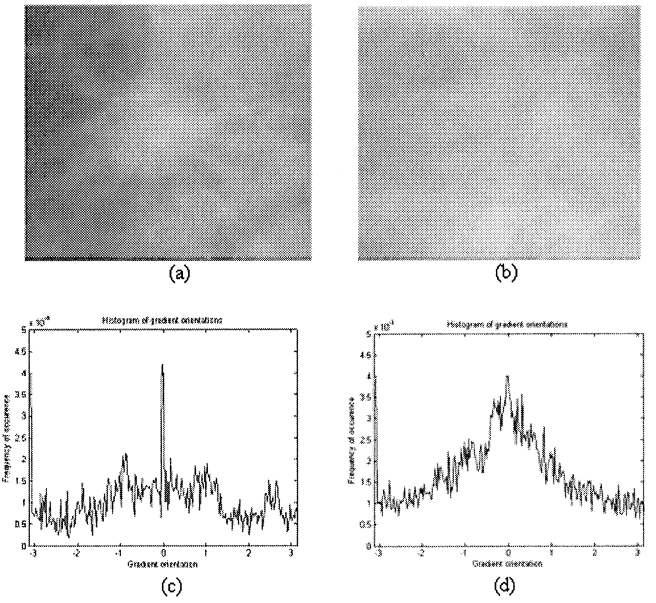

Fig. 2. Regions of a spiculated mass (B-3390_LCC) (a) and a normal dense tissue (C-0166_LCC) (b) as well as their gradient orientation histograms respectively (c and d).

Figure 2 provides two regions of mammograms, one with a spiculated mass (fig. 2a) and one with normal tissue (fig. 2b), along with their corresponding normalized gradient-orientation histograms (fig. 2c, 2d). The directions of spicules of the mass

differ from the directions of normal tissue. Specifically, pixels at normal areas have gradient orientations pointing to a certain direction range, while pixels at spiculated masses tend to have gradient orientations distributed in more directions. Therefore, the standard deviation of gradient orientations in the neighborhood of a mass pixel will be larger than that in the neighborhood of a normal pixel. As a result, the gradient orientation histogram is flat near a mass area (fig. 2c), but has a peak in areas containing normal dense tissue (fig. 2d).

2.4 Mass vs. Normal Tissue Classification

Logistic regression analysis was employed to determine the optimal subset of features that can differentiate masses from normal dense tissue. Specifically, feature-based logistic regression model was constructed by means of the forward feature selection method based on the area under ROC curve (Az) as a feature performance metric. The feature with the best Az value is first entered in the logistic regression model. The second feature selected is the one that in conjunction with the first feature yields the highest Az value among the remaining features. This process continues until no significant increase in terms of Az value is offered by adding features. After logistic regression model construction the half-half training and testing methodology was applied. The performance of the logistic regression model was evaluated in terms of Az area and standard error.

To study the effect of mass type (spiculated and circumscribed) in classification accuracy, logistic regression models were constructed for three differentiation tasks: (a) spiculated masses from normal dense tissue (S-N), (b) circumscribed masses from normal dense tissue (C-N) and (c) both spiculated and circumscribed masses from normal dense tissue (B-N).

3 Results

The ROC curves produced from the feature-based logistic regression models for the three differentiation tasks are presented in Figure 3. The Az values are 0.895 0.036, 0.875 0.033 and 0.813 0.032 for the C-N, S-N and the B-N datasets, respectively.

The differences in Az values between C-N and B-N, as well as between S-N and B-N are statistically significant (two-tailed student's t-test, $p<0.05$), indicating the reduction in classification accuracy when all masses (spiculated and circumscribed) are considered in the differentiation task.

4 Discussion and Conclusion

Our preliminary results suggest that histogram, texture and orientation-based features extracted from the coefficients of an overcomplete wavelet transform in combination with logistic regression analysis can provide a successful classification scheme for the detection of spiculated and circumscribed masses in dense parenchyma, as proved by ROC analysis. The most discriminating features seem to be the Skewness, Standard Deviation of Gradient-Orientation and Standard Deviation of Folded

Gradient-Orientation. On the other hand, Laws' texture measures do not possess any significant information, although they have been used in similar classification tasks [5].

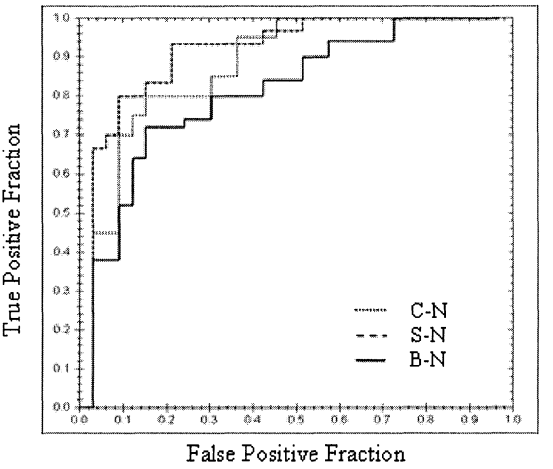

Fig. 3. The ROC curves for mass classification in the three differentiation tasks.

The performances (Az values) achieved in the mass classification accuracy are higher than those reported in other studies for dense breast category [13,14]. These encouraging results are in support of further development of the proposed method into a fully automated mass detection method. Future efforts will focus on: (a) extraction of additional features (e.g. coherence, entropy), (b) use of other classification techniques and (c) validation using a larger dataset.

5 Acknowledgements

We would like to thank the European Social Fund (ESF), Operational Program for Educational and Vocational Training II (EPEAEK II), and particularly the Program PYTHAGORAS I (B.365.011), for funding the above work.

References

1. R.T., Greenlee, M.B. Hill-Harmon, T. Murray, M. Thun, Cancer statistics CA. Cancer J. Clin. 51, 15-36 (2001).
2. J. Heine, P. Malhotra, Mammographic tissue, breast cancer risk, serial image analysis and digital mammography, Acad. Radiol. 9, 298-316 (2002).
3. L. Costaridou, S. Skiadopoulos, P. Sakellaropoulos, C.P. Kalogeropoulou, E. Likaki, G. Panayiotakis, Evaluating the effect of a wavelet enhancement method in characterization of simulated lesions embedded in dense breast parenchyma, Eur. Radiol. 15, 1615-1622 (2005).

4. W. Dahnert, Breast, in: Radiology Review Manual, edited by W. Dahnert (Williams and Wilkins, Baltimore, 1996), pp. 402-418.
5. J. Kegelmeyer, J.M. Pruneda, P.D. Bourland, A. Hillis, M.W. Riggs, M.L. Nipper, Computer-aided mammographic screening for spiculated lesions, Radiology 191, 331–337 (1994).
6. D. Wei, H-P. Chan, M.A. Helvie, B. Sahiner, N. Petrick, D.D. Adler, M.M. Goodsitt, Classification of mass and normal breast tissue on digital mammograms: Multiresolution texture analysis, Med. Phys. 22, 1501-1513 (1995).
7. D. Wei, H-P. Chan, N. Petrick, B. Sahiner, M.A. Helvie, D.D. Adler, M.M. Goodsitt, False-positive reduction technique for detection of masses on digital mammograms: Global and local multiresolution texture analysis, Med. Phys. 24, 903-914 (1997).
8. S. Liu, C.F. Babbs, E.J. Delp, Multiresolution detection of spiculated lesions in digital mammograms, IEEE Trans. Image Proc. 10, 874-884 (2001).
9. N. Petrick, H-P. Chan, D. Wei, B. Sahiner, M.A. Helvie, D.D. Adler, Automated detection of breast masses on mammograms using adaptive contrast enhancement and texture classification, Med. Phys. 23, 1685-1696 (1996).
10. H. Kobatake, M. Murakami, H. Takeo, S. Nawano, Computerized detection of malignant tumors on digital mammograms, IEEE Trans. Med. Imaging 18, 369-378 (1999).
11. R. Zwiggelaar, T.C. Parr, J.E. Schumm, I.W. Hutt, C.J. Taylor, S.M. Astley, C.R.M. Boggis, Model-based detection of spiculated lesions in mammograms, Med. Im. Anal. 3, 39-62 (1999).
12. I. Christoyianni, A. Koutras, E. Dermatas, G. Kokkinakis, Computer aided diagnosis of breast cancers in digitized mammograms, Comp. Med. Im. Graph. 26, 309-319 (2002).
13. W.T. Ho, P.W.T. Lam, Clinical performance of computer-assisted detection (CAD) system in detecting carcinoma in breast of different densities, Clin. Radiol. 58, 133-136 (2003).
14. G.D. Tourassi, R. Vargas-Voracek, D.M. Catarious, C.E. Floyd, Computer-assisted detection of mammographic masses: A template matching scheme based on mutual information, Med. Phys. 30, 2123-2130 (2003).
15. M. Heath, K. Bowyer, D. Kopans, R. Moore, R. Kegelmeyer, The digital database for screening mammography. in: Proceedings of 5th International Workshop on Digital Mammography, edited by M.J. Yaffe (Medical Physics Publishing, Madison, WI, 2000), pp. 212-218.
16. P. Sakellaropoulos, L. Costaridou, G. Panayiotakis, An image visualization tool in mammography, Med. Inform. 24, 53-73 (1999).
17. P. Sakellaropoulos, L. Costaridou, G. Panayiotakis, A wavelet-based spatially adaptive method for mammographic contrast enhancement, Phys. Med. Biol. 48, 787-803 (2003).
18. L. Costaridou, P. Sakellaropoulos, S. Skiadopoulos, G. Panayiotakis, Locally adaptive wavelet contrast enhancement. In: Medical Image Analysis Methods, edited by L. Costaridou (Taylor & Francis Group LCC, CRC Press: Boca Raton, FL, 2005), pp. 225-270.
19. L. Costaridou, P. Sakellaropoulos, A. Stefanoyiannis, E. Ungureanu, G. Panayiotakis, Quantifying image quality at breast periphery vs. mammary gland in mammography using wavelet analysis, Br. J. Radiol. 74, 913-919 (2001).