Automated Mammogram Classification Using a Multiresolution Pattern Recognition Approach

CRISTIANE BASTOS ROCHA FERREIRA
DÍBIO LEANDRO BORGES

PPGIA – Programa de Pós-graduação em Informática Aplicada
PUCPR – Pontifícia Universidade Católica do Paraná, R. Imaculada da Conceição, 1155, Prado Velho, 80215-901, Curitiba, PR, Brazil
{cristiane,dibio}@ppgia.pucpr.br

Abstract: In order to fully achieve automated mammogram analysis one has to tackle two problems: classification of radial, circumscribed, microcalcifications, and normal samples; and classification of benign, malign, and normal ones. How to extract and select the best features from the images for classification is a very difficult task, since all of those classes are basically irregular textures with a wide visual variety inside each class. In this paper we propose a multiresolution pattern recognition approach for this problem, by transforming the data of the images in a wavelet basis, and then using special sets of the coefficients as the features tailored towards separating each of those classes. For the experiments we have used samples of images labeled by physicians. Results shown are very promising, and the paper describes possible lines for future directions.

1 Introduction

Feature selection and classification are the cornerstone processes of a pattern recognition problem. In the case of image data one has to decide whether arrangements of spatial data (i.e. pixels directly) can be used as elements of features, or if a transformation of the pixels to a different space can uncorrelate the meaningful information needed to separate the data into the classes desired. For image analysis problems a texture is a challenging feature to recognize, since it is often not a regular pattern and it is very dependent on scale.

Breast cancer is the second major occurrence of cancer in Brazil, for women over 40 years old, being detected in more than 28,000 women in 2000, Kligerman [5]. An early diagnostic is very important, and one common method of diagnosis is by using a mammogram, which is basically an x-ray of the breast region taken in a special condition. Figure 1 shows a typical mammogram. From the image a trained physician screens it searching for microcalcifications, and masses which can be spiculated or circumscribed. If found, these artefacts on the image could be a sign for the presence of a benign or malign tumor. This is yet a challenging and unsolved problem for typical pattern recognition approaches, mainly because the microcalcifications and masses appear as almost free shapes, and there are vessels and muscles which are more or less prominent in the images depending on the patient.

We propose in this paper a novel multiresolution approach to perform an automated classification of mammograms. The experiments performed show that a successful classification can be achieved, even when we consider the two main problems: 1) Classification between normal, benign, and malign areas; 2) Classification between normal, microcalcifications, radial or spiculated, and circumscribed areas.

Section 2 shows the images of typical mammograms and its target classes, along with a revision of literature on mammograms classification. Section 3 defines the problem in terms of a pattern recognition framework and presents a proposed multiresolution approach for its solution. Section 4 shows experiments on images taken from MIAS [2]. Section 5 gives conclusions and points to future extensions.

2 Mammograms

The acquisition of a mammogram is done by compressing the breast of the patient between two acrylic plates for a few seconds when the x-ray is emitted.

Thus a typical mammogram is an intensity image with gray levels, showing the levels of contrast inside the breast which characterize normal tissue, vessels, different masses of calcification, and of course noise. An example of a mammogram is shown in Figure 1.

1530-1834/01 $10.00 © 2001 IEEE

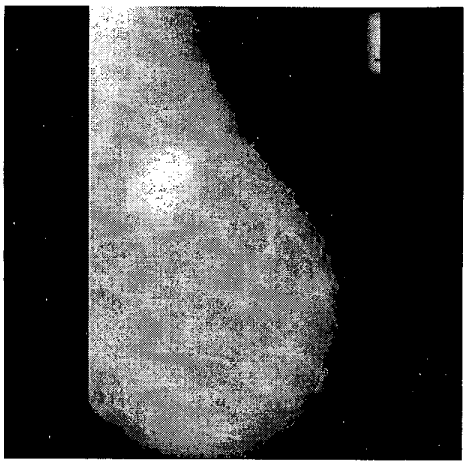

Figure 1: Intensity image of a typical mammogram (mdb184).

Some calcifications can be grouped in classes due their similar geometrical properties. They are usually named as radial or spiculated lesions, circumscribed masses lesions and microcalcifications.

The radial lesions have a centred region with segments leaving it in many directions, typical images are shown in Figure 2. The circumscribed masses lesions are more uniform, resembling a circle, although still irregular. Some examples are shown in Figure 3. Finally, the microcalcifications constitute small groups of calcified cells without pre-defined form or size. Figure 4 shows examples of microcalcifications.

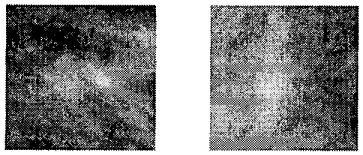

Figure 2: Typical radial lesions (mdb148 and mdb145, respectively).

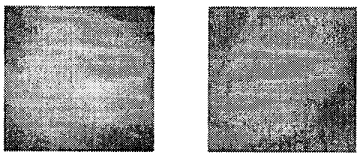

Figure 3: Typical circumscribed masses lesions (mdb028 and mdb019, respectively).

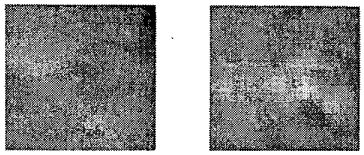

Figure 4: Examples of microcalcifications (mdb238 and mdb248, respectively).

Another classification adopted by a physician considers the nature of the lesions, such as benign or malign lesions.

The distinction between these two classes is very ill defined in terms of the images themselves, since what usually a physician does is to ask for further analysis including other tests for characterizing the tumor as benign or malign. In terms of an automated classification to be performed by computer, a strong evidence of a classification in one of these classes will be an important aimed result. In our approach here we will use typical examples of images in the classes in order to characterize the class. The images used in the experiments were labeled by a physician, MIAS [2].

Mammograms without any of the typical artefacts, or abnormalities will be classified as normal cases. Typical images of this class are shown in Figure 5.

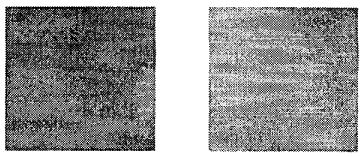

Figure 5: Examples of normal cases (mdb033 and mdb037, respectively).

The mammograms used in this work came from the Minimammographic database of MIAS – Mammographic Image Analysis Society [2], with original size of 1024x1024 pixels, per image, and namely mdbXXX, where XXX will be a number of the image in the database. However, the images used in the experiments were cuttings of size 64x64 pixels done in the original mammograms, whose centers correspond to the centers of the presented abnormalities.

A glance through the images in Figures 2, 3, 4, and 5 shows that one can not deal with this type of image analysis problem by looking for a well behaved visual template, or structural data. The images are irregular textures, and with subtle similarities and differences regarding the classification between radial, circumscribed, microcalcifications, and normal; or between normal, benign, and malign.

A solution to this whole problem is still a research issue. Some works from the literature either deal only

with the segmentation of mammograms in order improve visualization and analysis by a physician, or classify subsets of classes. A review of some work until 1994 can be seen in Woods [12]. We will comment here on some recent works.

Rangayyan et al. [10] present a scheme for analysing mammograms by using a multiresolution representation based on Gabor wavelets. The method is used to detect asymmetry in the fibro-glandular discs of left and right mammograms in order to diagnose breast cancer. The types of lesions are not dealt with as it is the approach taken here. In their work a dictionary of Gabor filters is used and the filter responses for different scales and orientation are analysed by using the Karhunen-Loève transform, which is applied to select the principal components of the filter responses. They show figures of correct classification for asymmetric, distortion, and normal cases.

Guimarães [1] proposes the use neural networks and fuzzy logic for classifying benign, and malign tumors. The images were pre-segmented and featured as rugosity vectors, which are then trained and passed for defuzzification rules for ranking. In this case the separation between the two classes was facilitated by the high degree of rugosity found in the samples used for testing the two classes.

A more similar approach to ours is one by Pereira [9], which tries to classify mammograms into normal, benign, malign, and inside the last two classes between types of lesions such as radial, circumscribed, and microcalcifications. In Pereira [9] a multiresolution representation of each mammogram is computed using Haar, Daubechies, and a shiftable transform. The classification is attempted by using a metric proposed by Jacobs, Finkelstein and Salesin [3]. Some results are given showing successful rates.

Our approach here differs from those by first devising a multiresolution representation of a mammogram class which is based on the statistical properties of a wavelet transform, and then denoise the coefficients in order to achieve a separable nearest neighbor classifier for the two problems.

Section 3 next frames the problem in a pattern recognition framework and presents the details of our approach.

3 Texture analysis and a pattern recognition framework using a multiresolution approach

In a general way texture can be characterized as the space distribution of the gray levels in a neighborhood, as in Jain, Kasturi, and Schunck [4], that is to say, the variation pattern of the gray levels in a certain area.

Texture is a feature that can not be defined for a point, and the resolution at which an image is observed determines the scale at which the texture is perceived. So, texture is a confusion measurement that depends mainly on the scale which the data are observed.

There are textures with regularity, deterministic and structured aspects, and others irregular like the mammograms previously shown.

In case of regular textures, some measurements can be used like gray-level co-occurrence matrices to capture the spatial dependence of gray-levels values. In addition, entropy, energy, contrast and homogeneity properties can be calculated easily. An autocorrelation function also can be used for images with repetitive texture patterns because it exhibits periodic behavior with a period equal to spacing between adjacent texture primitives.

However, in our problem, the images are mammograms with irregular textures, and in addition, the mammogram classes are not homogeneous. Therefore, those measurements will not be representative for the kind of classes we aim to separate in an automated mammogram analysis.

We need first to find what features can be useful, and then select possibly uncorrelated measurements of them. As proposed here this can be achieved by using a wavelet transformation of the data, since the statistical properties of these types of transforms will help us to untangle the data.

Since we are dealing with images with not regular textures, the elements of each class are more separated in the feature space. The problem requires a transform which can uncorrelate the data as much as possible without losing their main distinguishable characteristics. A wavelet transform, Vidakovic [11], will be useful here, since it has the property of uncorrelate the data, besides its processing algorithm is O(n), and for the problem at hand one can denoise the data in order to achieve a further dimensionality reduction. The main contribution of our approach presented here is to derive such an approach suited to the problem of automating mammogram analysis.

Wavelets are applied in several areas, such as signal processing, temporal series analysis, meteorology, image filtering and compression, and pattern identification. In our work we have used wavelet transform to propose a pattern recognition solution for mammogram analysis.

The wavelets are functions used as basis for representing other functions, and once a so called mother wavelet is fixed, a family can be generated by translations and dilations of it. If we denote a mother wavelet as $\psi(x)$, its dilations and translations are

$$\{\psi(\frac{x-b}{a}),(a,b)\in R^{+}\times R\} \quad (1)$$

where $a = 2^{-j}$ and $b = k \cdot 2^{-j}$, with k and j integers.

The wavelets used in the experiments of this work were Haar and Daubechies 4 (Db4), Meyer [8], implemented following the multiresolution scheme given by Mallat [6].

A bidimensional wavelet can be understood as an unidimensional one along axes x and y. In this way applying convolution of low and high pass filters on the original data, the signal can be decomposed in specific sets of coefficients, at each level of decomposition, as:

low frequency coefficients ($A^d_{2^j}f$),

vertical high frequency coefficients ($D^1_{2^j}f$),

horizontal high frequency coefficients ($D^2_{2^j}f$), and

high frequency coefficients in both directions ($D^3_{2^j}f$).

The $A^d_{2^j}f$ coefficients represent the entry of next level of decomposition. The decomposition process proposed by Mallat [6] and implemented in our work represents the pyramidal algorithm for a bidimensional wavelet transform. Figure 6 shows a diagram of the decomposition process.

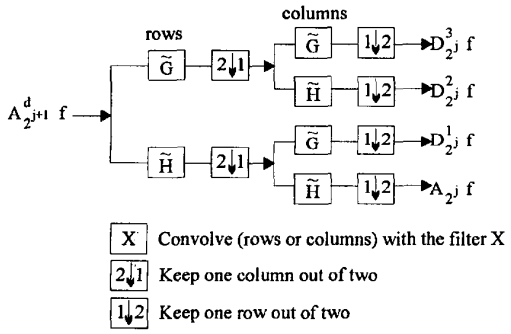

Figure 6: Decomposition process for computing a wavelet transform.

After obtaining the four sets of coefficients it is necessary to establish what, and how many coefficients have the capability for sufficient and satisfactory representation of texture, extracting relevant information to represent the original mammogram, and to aid in the classification process.

In our work, we propose that it is not necessary to keep and use all of the coefficients for classification. In fact, to our knowledge this is the first attempt to devise thresholding strategies to suit a pattern recognition problem, instead of using the coefficients for compression and noise supression. We propose and test keeping only the 100 greatest coefficients $A^d_{2^j}f$ in magnitude of the decomposed image in first level of decomposition. We have run experiments using Daubechies Db4 and Haar wavelets. Each class was represented for normalized mean of the 100 greatest coefficients $A^d_{2^j}f$.

The mammograms used for tests were decomposed for the same decomposition process, retaining the same amount of coefficients mentioned previously.

Thus, a nearest neighbor classifier is designed using euclidean distance as a metric between the correspondent normalized wavelet coefficients, as shown in equation 2. A prototype of each class is first derived using a mean vector of labeled images, then tests are run for all of the other images from the database.

$$D_{Euclidiana} = \sqrt{\sum_{i,j} (A(i,j) - M(i,j))^2} \qquad (2)$$

where A is a matrix of selected wavelet coefficients, M is the prototype of a class, and the distance is computed only for those terms where $A(i,j) \neq 0$.

Therefore, A will belong to the class which has the minimum distance.

4 Experiments

Experiments were accomplished for the two problems: the geometric property of the tumor, and its nature.

The first set of experiments took into consideration the geometric property of the tumor, considering four classes: radial lesions, circumscribed lesions, microcalcifications and normal areas.

The images used in this set of experiments are shown by class. Some noisy images were obtained from original ones and used for testing, namely ndbXXX, rdbXXX and sdbXXX.

The noisy images were obtained by application of three types of noise, with GIMP [7] image processor: Noisify, Randomize and Spread, corresponding to ndbXXX, rdbXXX and sdbXXX, respectively. The parameter settings were independent option and gray factor equals 10 to Noisify. In case of Randomize, randomization type was pick, randomization seed was 1, randomization percentile was 100% and 10 number of repetitions. At last, in case of Spread, both horizontal and vertical spread amount were 10.00.

The images used for constructing the classes are different from the images used for classification. Figures 7, 8, 9 and 10 show images of radial, circumscribed lesions, microcalcifications, and normals respectively.

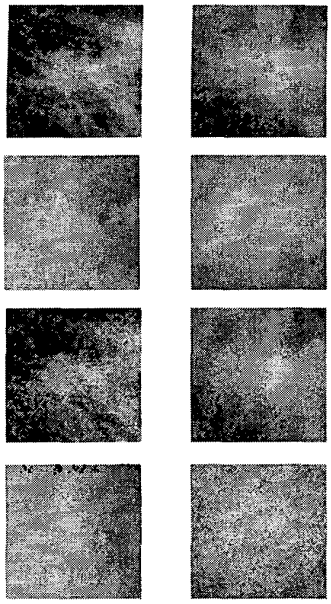

Figure 7: Radial lesion class
(mdb148, mdb181, mdb145, mdb199, ndb148, sdb181, rdb145 and ndb199, from left to right, from top to bottom, respectively).

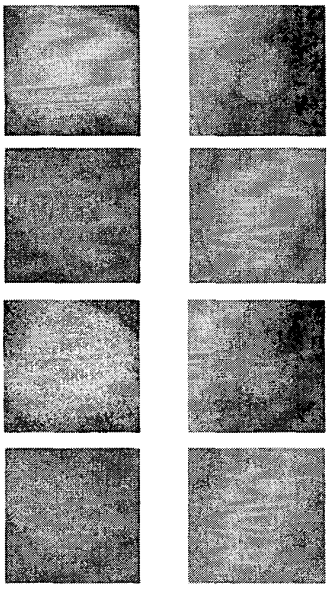

Figure 8: Circumscribed lesion class
(mdb028, mdb270, mdb012, mdb019, ndb028, sdb270, rdb012 and sdb019, from left to right, from top to bottom, respectively).

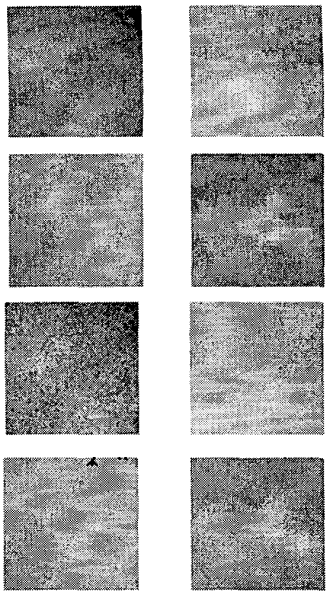

Figure 9: Microcalcification class
(mdb238, mdb253, mdb227, mdb248, ndb238, sdb253, rdb227 and sdb248, from left to right, from top to bottom, respectively).

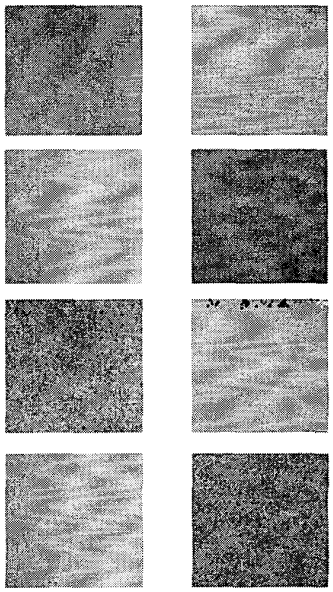

Figure 10: Normal class
(mdb033, mdb037, mdb004, mdb070, ndb033, rdb037, sdb004 and ndb070, from left to right, from top to bottom, respectively).

The Db4 and Haar wavelets were the basis used in the decomposition process and the 100 greatest coefficients $A^d_2 f$ in magnitude in the first level of decomposition were considered.

Table 1 shows the successful rates concerning the results obtained with this set of experiments. In this case, using either of the wavelet basis, the results were similar. All of the mammograms belonging to the microcalcification and normal classes were correctly classified, scoring 100.0%. For radial and circumscribed lesions the figure of correct classification was 91.7%. Here, some images were confused between radial and circumscribed lesions.

Class	**Wavelet**	
	Db4	Haar
Radial lesions	91.7	91.7
Circumscribed lesions	91.7	91.7
Microcalcification	100.0	100.0
Normal	100.0	100.0

Table 1: Successful rates of classification, in percentage, for the first set of experiments.

The second set of experiments took into consideration the nature of the tumor, regardless of geometric property, considering three classes: benign, malign and normal.

Figures 11, 12 and 13 show benign, malign and normal classes respectively.

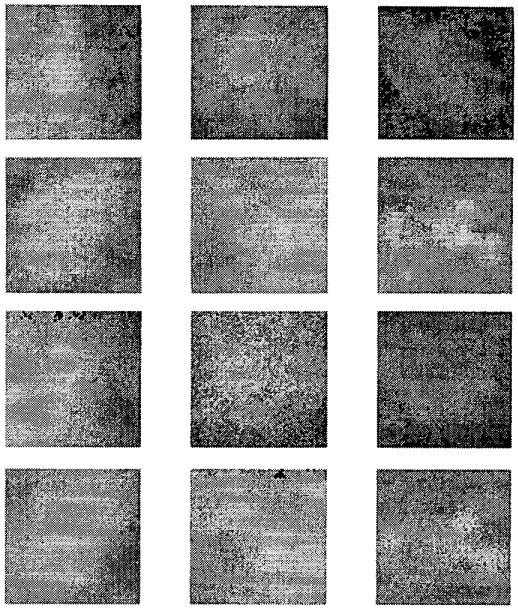

Figure 11: Benign class (mdb145, mdb199, mdb012, mdb019, mdb227, mdb248, rdb145, ndb199, rdb012, sdb019, rdb227 and sdb248, from left to right, from top to bottom, respectively).

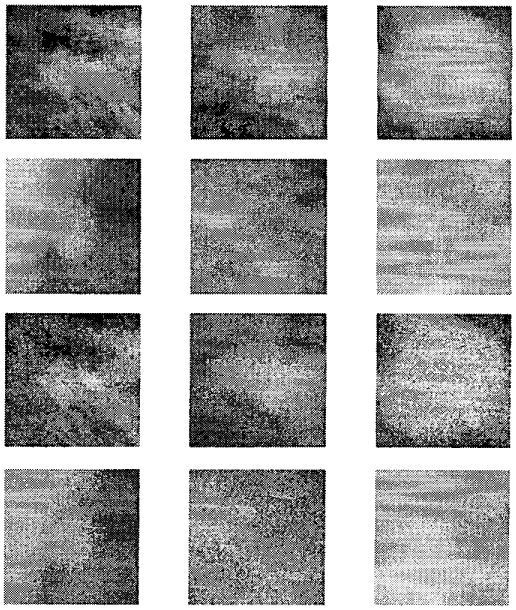

Figure 12: Malign class (mdb148, mdb181, mdb028, mdb270, mdb238, mdb253, ndb148, sdb181, ndb028, sdb270, ndb238 and sdb253, from left to right, from top to bottom, respectively).

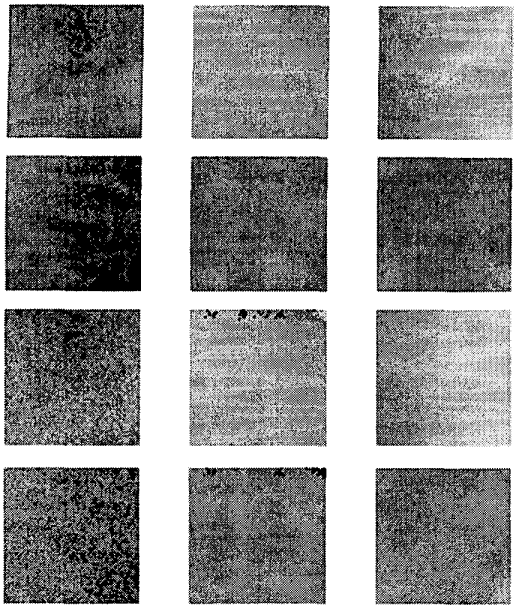

Figure 13: Normal class
(mdb033, mdb037, mdb004, mdb070, mdb024, mdb074, ndb033, rdb037, sdb004, ndb070, rdb024 and sdb074, from left to right, from top to bottom, respectively).

The Db4 and Haar wavelets were used again in the decomposition process and the 100 greatest coefficients $A^{d}{}_{2}{}^{j}f$ in magnitude in the first level of decomposition were considered.

Table 2 shows the successful rates of the classification for the second set of experiments. All of the benign images were classified correctly in both basis. For the malign class the successful rate achieved was 83.3%, also for Db4 and Haar wavelets. Finally, all of the normal images were classified correctly using Haar wavelet, but in case of Db4 wavelet, some images were confused with the malign class, resulting in a rate of 66.7% of correct classification.

Class	Wavelet	
	Db4	Haar
Benign	100.0	100.0
Malign	83.3	83.3
Normal	66.7	100.0

Table 2: Successful rates of classification, in percentage, for the second set of experiments.

The results shown here are promising, since solving the two problems, i.e. classification regarding geometric properties (Table 1), and classification regarding the nature of the tumor (Table 2). It is a very complex pattern recognition problem. Other wavelet basis can be tried, or even best designed for the problem, and results can even be improved. In this work we have shown that a special set of the wavelet coefficients can be used indeed for achieving a successful classification. We are not aware of similar results from the literature.

5 Conclusions

Having a complete automated mammogram analysis solution is still something to be achieved. In this paper we have shown that considering the two main classification problems, and by the samples of real mammograms, the solution has to tackle the irregular textures, and the large variation inside each class.

We have proposed in this paper that by using a wavelet transformation of the data, we can devise a pattern recognition solution where the features are selected directly from the wavelet decomposition, denoised in a tailored way to accomplish a separation between the classes, and processed in O(n) time, instead of more computational demanding algorithms such as KLT (Karhunen-Loève Transform) for example, Jain, Kasturi and Schunck [4]. In fact we have designed an special algorithm for that, and the results shown in this paper are very promising compared to others from the literature.

Future extensions of this approach will deal with a way to incorporate possibly others coefficients of the transform, either from high frequency decomposition, or other levels, since some useful distinct information can be found from those.

Acknowledgments

This work has been partially supported by PUCPR with a student bursary given to the first author C.B.R.Ferreira.

References

[1] M. S. Guimarães, *"Abordagens Difuso-Neurais para Análise de Mamogramas"*, M.Sc. dissertation, Centro Tecnológico de Computação Aplicada & Automação, Universidade Federal Fluminense, Niterói, Rio de Janeiro, Brasil, 1999.

[2] http://www.wiau.man.ac.uk/services/MIAS (Mammographic Image Analysis Society).

[3] C. E. Jacobs, A. Finkelstein and D. H. Salesin, "Fast Multiresolution Image Querying", *Proceedings of SIGGRAPH 95*, (1995), 227--286.

[4] R. Jain, R. Kasturi and B. Schunck, *"Machine Vision"*, McGraw Hill, USA, 1995.

[5] J. Kligerman, "Estimativas sobre a Incidência e Mortalidade por Câncer no Brasil - 2000", *Revista Brasileira de Cancerologia*, vol. 46, n. 2, (2000), 1--5.

[6] S. G. Mallat, "A Theory for Multiresolution Signal Decomposition: The Wavelet Representation", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, n. 7, (July 1989), 674--693.

[7] P. Mattis, S. Kimball. "The GNU Image Manipulation Program", released the GNU General Public License (GPL), for UNIX and X. version 0.99.23. (http://www.gimp.org)

[8] Y. Meyer, *"Wavelets and Operators"*, Cambridge Univ. Press, England, 1992.

[9] R. C. Pereira, *"Identificação de Tumores em Mamogramas através de Representações Wavelets"*, M.Sc. dissertation, Escola de Engenharia Elétrica, Universidade Federal de Goiás, Goiânia, Goiás, Brasil, 1999.

[10] R. M. Rangayyan, R. J. Ferrari, J. E. L. Desautels and A. F. Frère, "Directional Analysis of Images with Gabor Wavelets", *Proceedings of XIII Brazilian Symposium on Computer Graphics and Image Processing, SIBGRAPI 2000,* (2000), 170--177.

[11] B. Vidakovic, "Nonlinear Wavelet Shrinkage with Bayes Rules and Bayes Factors", *J. Amer. Statistical Association*, vol. 93, n. 441, (1998), 173--179.

[12] K. S. Woods, *"Automated Image Analysis Techniques for Digital Mammography"*, Ph.D. Thesis, Dept. C. Science and Engineering, Univ. South Florida, FL, USA, 1994.