An Evaluation of Wavelet Features Subsets for Mammogram Classification

Cristiane Bastos Rocha Ferreira[1] and Díbio Leandro Borges[2]

[1] Universidade Federal de Goiás,
Instituto de Informática, Goiânia, Go, Brazil
cristiane@inf.ufg.br
[2] BIOSOLO, Goiânia, Go, Brazil
dibio.borges@terra.com.br

Abstract. This paper is about an evaluation for a feature selection strategy for mammogram classification. An earlier solution to this problem is revisited, which constructed a supervised classifier for two problems in mammogram classification: tumor nature, and tumor geometric type. The approach works by transforming the data of the images in a wavelet basis and by using a minimum subset of representative features of these textures based in a specific threshold (λ_T). In this paper different wavelet bases, variation of the selection strategy for the coefficients, and different metrics are all evaluated with known labelled images. This is a suitable solution worth further exploration. For the experiments we have used samples of images labeled by physicians. Results shown are promising, and we describe possible lines for future directions.

1 Introduction

In a pattern recognition approach, the features used to represent the classes must be significative to characterize them with precision and to contribute positively towards the classification process. In the case of images, a transformation of pixels to a different space can help to untangle the meaningful information.

An early diagnostic for medical treatment is very important to total or partial cure. This can avoid the surgical removal of a breast. A common method of diagnosis is by using a Mammogram, which is basically an x-ray of the breast region that displays points with bigger intensities. From the image a trained physician screens it searching for artifacts that could be a sign for the presence of a benign or malign tumor. However suspicious areas appear as almost free shapes and this a challenging for pattern recognition approaches. Besides there are vessels and muscles which are more or less prominent in the images depending on the patient. The variation of images in a class and among considered classes is a factor that will influence directly the problem treated in this paper.

We proposed a solution to this in a previous paper [3] using feature sets with 100, 200, 300 and 500 features to represent each image class. In this paper we report on an strategy to select the wavelet features to be used n the classification, and further it is shown a protocol of tests evaluating the features chosen on two

M. Lazo and A. Sanfeliu (Eds.): CIARP 2005, LNCS 3773, pp. 620–630, 2005.

mammogram classification problems: 1) Type(Benign or Malign) and presence of tumor; and 2) Shape of the Artifacts Distribution in he Mammogram. Considerable advances in this paper are achieved if compared with [3], because of the reduction of the dimensions in space and successful classification rates. This reduction is provided by a new strategy to select the most significant features based on standard deviation of classes by a specific threshold λ_T.

A mammogram classifier is constructed and evaluated using a wavelet decomposition process and a selected subset of representative features. The experiments performed show that 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 approach for its solution. Section 4 shows experiments on images taken from MIAS [4]. Section 5 gives conclusions and points to future extensions.

2 Mammograms

A mammogram is an x-ray of breast obtained by compression of the breast of patients between two acrylic plates for a few seconds. 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. This type of image is used by physicians because it is cheap and it allows the discovery of breast cancer that is not perceived in a touch verification. An example of a mammogram and a machine used for obtaining this type of image are shown in Figure 1 a) and b), respectively.

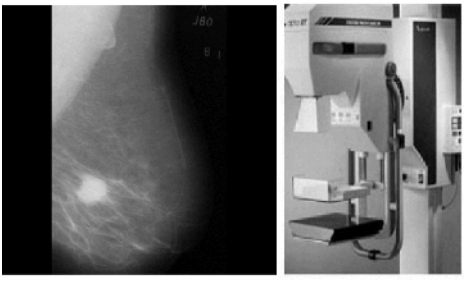
Some calcifications can be grouped in classes due their similar geometrical properties. They are usually named radial or spiculated lesions, circumscribed masses lesions and microcalcifications. The radial lesions have a centred region with segments leaving it in many directions. The circumscribed masses lesions are more uniform, resembling a circle, although still irregular. Finally, the microcalcifications constitute small groups of calcified cells without pre-defined form or size.

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 a computer, a strong evidence of a classification in one of these classes will be an important result to achieve. Mammograms without any of the typical artifacts, or abnormalities will be classified as normal cases.

The images used in the experiments were labelled by a physician and they came from the database of MIAS [4] with original size of 1024x1024 pixels, per

image, and namely mdbX, where X is a number of the image in the database. However, the images used in the experiments were crops of size 64x64 pixels performed in the original mammograms, whose centers correspond to the centers of the presented abnormalities. 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. Figures 5 and 6 show examples of the two classification problems addressed here.

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 [9]. We will comment here on some recent works.



a) b)

Fig. 1. a) Intensity image of a typical mammogram (mdb184) b) Mammogram machine

In [8] is presented a scheme for analyzing 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 analyzed 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. In [7], thermal texture maps are used in early detection of breast cancer. In this case the relationship between the pattern in each slice and the metabolic activities within a patient's body is revealed and the depth of tumor is estimated by thermal-electric analog and half power point. The conclusion is based on fact that different tissues have different growth patterns and this can distinguished the pixels of tumors and blood vessel. This approach is used to detection of breast cancer and ovarian cancer.

This paper represents the continuity of approach presented in [3] and it shows the constructing and evaluating of classifier for mammogram using a wavelet decomposition process for the feature extracting stage. We evaluate a different strategy for representative feature selecting is presented by using a specific

threshold (λ_T), based on standard deviation of classes. The number of features is reduced drastically and results shown have high successful rates.

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

3 The Proposed Approach

In a general way texture can be characterized as the space distribution of the gray levels in a neighborhood, as in [5], 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. This can be reached by using a wavelet transform in data, because statistical properties of this kind of transformation can help to uncorrelate the data as much as possible without losing their main distinguishable characteristics.

The main contribution of this method is the design and selection of a feature representation of mammogram that can help in the mammogram classification process. We use a wavelet transform in data and we reach a dimensionality reduction. We propose a selecting strategy of main features subsets that have a good representation for the elements of each class and they are more separated in the feature space. A specific threshold (λ_T) based on standard deviation of classes images is used. Extracted and selected features of the decomposed image are used in the construction of the image signature. We believe that this approach can be used in other applications that deal with recognition of irregular textures, like other medical image applications. In order to achieve a separation among image for experiments, the following conventions are adopted: "Basis Image" for mammogram subset with known classification and "Test Image" for mammogram subset with unknown classification, used in test stage.

3.1 Wavelets

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 y(x), its dilations and translations are

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

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

The wavelets used in the experiments of this work were implemented following the multiresolution scheme given by Mallat [6].

A bi-dimensional wavelet can be understood as an one-dimensional 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 ($A_2^d jf$);
- vertical high frequency coefficients ($D_2^1 jf$),
- horizontal high frequency coefficients ($D_2^2 jf$), and
- high frequency coefficients in both directions ($D_2^3 jf$).

The $A_2^d jf$ 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 bi-dimensional wavelet transform. Figure 2 represents the wavelet decomposition process and Figure 3 show an example of decomposed mammogram.

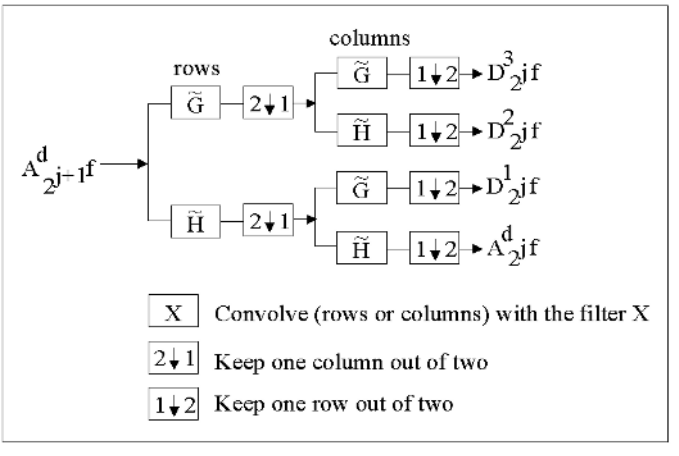


Fig. 2. Decomposition process for computing a wavelet transform

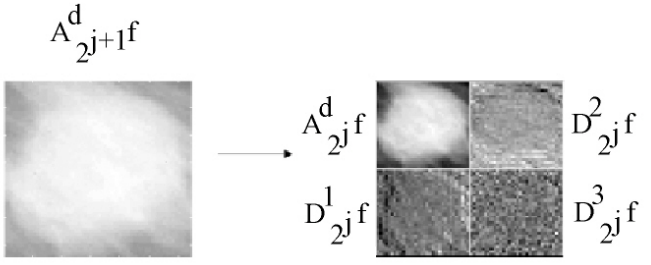


Fig. 3. Example of a decomposed mammogram

By having in mind that decomposing the input image with a Wavelet Transform will be a pre-processing step, the approach can be described then in two main stages as follows.

3.2 First Stage: Building the Classes Signatures

The first stage is based on the Basis Image subset and it is based on the following steps:

- Mammogram images are decomposed with a chosen wavelet basis (W_i);
- Some low frequency coefficients ($CoefClass_j$) are selected, based on their magnitude, in the first decomposition level, considering λ_T as the threshold;
- Signatures of the classes ($ClassSig_j$) are built based on $CoefClass_j$ and on the mean of those coefficients.

The λ_T threshold is calculated using λ [2] defined as:

$\lambda = \frac{\sigma\sqrt{2logn}}{n}$

where σ represents the standard deviation of the class and n represents the number of images in that class.

The λ_T threshold is calculated by a mean of the λ thresholds of j classes, e. g.:

$\lambda_T = \frac{\sum_{v=1}^{j} \lambda_v}{j}$,

where j represents the number of classes considered.

3.3 Second Stage: Classifying a Mammogram

The second stage is based on the Test Image subset and follows the procedures presented below:

- An unknown mammogram ($Mamo_k$) is decomposed with a chosen wavelet basis (W_i);
- Some low frequency coefficients ($CoefMamo_k$) are selected, based on their magnitude, in the first decomposition level, considering λ_T as a threshold;
- In the second stage, $CoefClass_j$ coefficients represent the unknown mammogram signature ($MamoSig_k$)

- Distances between $MamoSig_k$ and $ClassSig_j$ signatures are calculated by different metrics. D_j are computed for all classes $ClassSig_j$;
- The unknown mammogram is classified based on the lowest distances D_j.

The distance metrics used in order to measure the proximity between unknown mammogram and classes signatures are: Euclidean Distance, Norm in Absolute Value, Mahalanobis Distance and Huffmann Code. The Euclidean Distance is defined by

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

The Norm in Absolute Value is represented by

$$D_{AbsoluteValue} = \sum_{i,j}(A(i,j) - M(i,j)).$$

A is the matrix that represents the mammogram signature ($MamoSig_k$), M is the class signature ($ClassSig_j$) and the distance is calculated for all $A(i,j) \neq 0$. Mahalanobis Distance is defined by

$$D^2_{Mahalanobis} = (x - m)'C^{-1}(x - m),$$

where x is the features' matrix of mammogram that it is to be classified represented by $MamoSig_k$, m is the matrix of arithmetic mean among all of elements of the same class, represented by $SigClass_j$, and C^{-1} is the covariance matrix of class elements. Huffmann Code is based on the following rules: for an A matrix, for all i and j, we have $A(i,j) = 1$, if $A(i,j) > 0$, and $A(i,j) = -1$, if $A(i,j) < 0$, where i is the number of lines and j is the number of columns of A matrix. Considering that A and B are matrices, the distance between them, using Huffmann Code is calculated by a sum of "1", where the sum is calculated in cases where $A(i,j) = B(i,j)$.

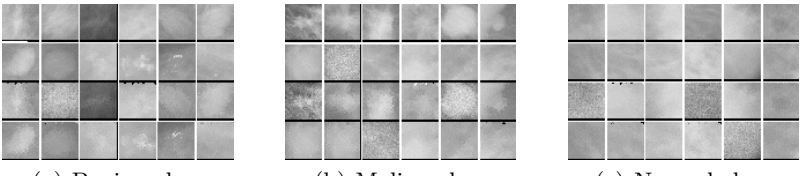
4 Experiments and Analysis of Results

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 second experiment took into consideration the nature of the tumors, regardless of geometric property, considering three classes: benign, malign and normal classes.

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 ndbX, rdbX and sdbX. The noisy images were obtained by application of three types of noise: Noisify, Randomize and Spread, corresponding to ndbX, rdbX and sdnX, respectively. The parameter settings were independent, option of gray factor equals 10 to Noisify. In case of Randomize, randomization percentile was 100% and 10 number of repetitions. At last, in case of Spread, both horizontal and vertical spread amount were 10.00.

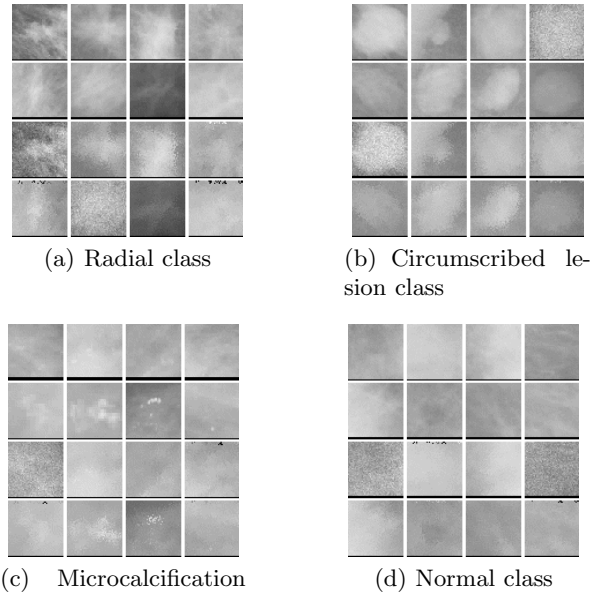
The images used for constructing the classes are different from the images used for classification. Figures 4(a), 4(b) and 4(c) show benign, malign and normal classes respectively, considering the nature of tumors. Figures 5(a), 5(b), 5(c) and 5(d) show radial lesions, circumscribed lesions, microcalcifications and normal classes, considering the geometric property of tumor.

We consider the variation of two issues: wavelet basis used in the decomposition process, and distance metrics. The wavelet bases tested were Haar, Daubechies 4, Biorthogonal 2.4, Coiflets 2 e Symlets 2. A cross validation process is performed with 75% of images separated for building the classes signatures and 25% of them for testing. Four rounds are tested with all of images considering the mentioned percentages and we present the average results in Tables 3, 4, 5, 6, and 7. Tables 1 and 2 present λ_T threshold values for each test, considering the nature and geometrical properties of tumors, respectively.



(a) Benign class (b) Malign class (c) Normal class

Fig. 4. Typical images of the classes for the first mammogram classification problem considered in this work (Tumor Nature)



(a) Radial class (b) Circumscribed lesion class

(c) Microcalcification class (d) Normal class

Fig. 5. Typical images for the second classification problem (Tumor Geometrical Type)

Table 1. λ_T values, considering nature of the tumors

Round	Daubechies 4	Haar	Biorthogonal 2.4	Coiflets 2	Symlets 2
1	15	15	14	14	15
2	13	14	12	12	14
3	15	15	14	14	15
4	14	14	12	12	14

Table 2. λ_T values, considering the geometrical properties of the tumors

Round	Daubechies 4	Haar	Biorthogonal 2.4	Coiflets 2	Symlets 2
1	15	15	14	14	15
2	17	17	15	15	17
3	15	15	14	14	15
4	17	17	16	16	17

Table 3. Successful rates of classification using Daubechies 4 wavelet basis with normalized data

Class	Euclidean Distance	Norm in Abs. Value	Huffmann Code	Mahalanobis Distance
Benign	95.83	95.83	87.50	95.83
Malign	45.83	33.33	45.83	33.33
Normal	87.50	83.33	87.50	87.50
Radial	56.25	50.00	56.25	50.00
Circumscribed	75.00	75.00	75.00	75.00
Microcalcifications	93.75	93.75	68.75	93.75
Normal	75.00	75.00	87.50	68.75

Table 4. Successful rates of classification using Haar wavelet basis with normalized data

Class	Euclidean Distance	Norm in Abs. Value	Huffmann Code	Mahalanobis Distance
Benign	91.67	91.67	91.87	91.67
Malign	79.17	70.83	79.17	66.67
Normal	95.83	95.83	100.00	95.83
Radial	75.00	75.00	75.00	75.00
Circumscribed	87.50	87.50	87.50	87.50
Microcalcifications	93.75	93.75	93.75	93.75
Normal	93.75	93.75	100.00	100.00

Table 5. Successful rates of classification using Biorthogonal 2.4 wavelet basis with normalized data

Class	Euclidean Distance	Norm in Abs. Value	Huffmann Code	Mahalanobis Distance
Benign	75.00	75.00	66.67	83.33
Malign	50.00	29.17	20.83	29.17
Normal	83.33	83.33	95.83	75.00
Radial	75.00	75.00	62.50	75.00
Circumscribed	62.50	62.50	62.50	62.50
Microcalcifications	75.00	62.50	62.50	56.25
Normal	62.50	56.25	81.25	56.25

Table 6. Successful rates of classification using Coiflets 2 wavelet basis with normalized data

Class	Euclidean Distance	Norm in Abs. Value	Huffmann Code	Mahalanobis Distance
Benign	87.50	87.50	70.83	87.50
Malign	83.33	58.33	54.17	54.17
Normal	87.50	83.33	95.83	83.33
Radial	81.25	81.25	81.25	81.25
Circumscribed	81.25	81.25	81.25	81.25
Microcalcifications	93.75	93.75	93.75	87.50
Normal	81.25	75.00	93.75	75.00

Table 7. Successful rates of classification using Symlets 2 wavelet basis with normalized data

Class	Euclidean Distance	Norm in Abs. Value	Huffmann Code	Mahalanobis Distance
Benign	87.50	87.50	83.33	91.67
Malign	79.17	70.83	75.00	54.17
Normal	95.83	95.83	100.00	95.83
Radial	68.75	62.50	75.00	62.50
Circumscribed	81.25	87.50	81.25	87.50
Microcalcifications	93.75	87.50	93.75	81.25
Normal	93.75	93.75	100.00	100.00

The experiments show that the distance metrics used in the classification process present similar results on average. Euclidean Distance and Norm in Absolute Value show similar successful rates, with the exception of some cases in the malign class. In some cases, Mahalanobis Distance presents inferior rates when compared to other metrics. Haar basis achieves better results considering all the tested classes. The dimensionality of feature space is reduced and the results are promising for the two mammogram classification problems. Selection of features by the λ_T threshold demonstrates its representation capability for choosing the minimum features subset used for building the signatures of classes. The number of features used is about of 1.46% of the low frequency coefficients in the first level of decomposition and 0.37% of total information. Thus relevant information is concentrated in few low frequency coefficients.

5 Conclusions and Future Works

This paper showed an evaluation of a feature selection strategy for two mammogram classification problems. We see this as a practical and important issue to be addressed in medical applications. Variations of the problem, considering tumor nature, and tumor geometric properties are considered. The strategy for the classification was first presented in [3], and in this work we have used a threshold, λ_T, to select the coefficients and have presented experiments in a different number of conditions. The λ_T threshold was capable to choose signatures that conduced to a representation that showed successful rates in classification

process, and with λ_T it was possible to use a smaller quantity of features that are useful for mammogram classification problem.

Future extensions of this approach will try to deploy a fully working system in a medical environment. In addition, we suggest the union of this process of decision making of classification with medical inference models of diagnosis.

References

1. P. A. Devijver, J. Kittler *Pattern Recognition: A Statistical Approach* Pretice-Hall, England (1982).
2. D. L. Donoho, I. M. JohnStone, "Ideal spatial adaptation by wavelet shrinkage", *Biometrika*, vol. 81, (1994), 425–455.
3. C. B. R. Ferreira, D. L. Borges, "Analysis of mammogram classification using a wavelet transform decomposition", *Pattern Recognition Letters*, 24, Holand (2003), 973-982.
4. http://www.wiau.man.ac.uk/services/MIAS (Mammographic Image Analysis Society).
5. R. Jain, R. Kasturi, B. Schunck *Machine Vision* McGraw Hill, USA (1995).
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. H. Qi, P. Kuruganti, Z. Liu, "Early detection of breast cancer using thermal texture maps", em IEEE Symposium on Biomedical Imaging: Macro to Nano, (2002)
8. R. M. Rangayyan, R. J. Ferrari, J. E. L. Desautels, A. F. Frère, "Directional analysis of images with Gabor wavelets", *In: Proceedings of XIII Brazilian Symposium on Computer Graphics and Image Processing, SIBGRAPI*, (2000), 170-177.
9. K. S. Woods, *Automated image analysis techniques for digital mammography*, Ph. D thesis, Dept C. Science and Engineering, University of South Florida, FL, USA (1994).