THERMAL FEATURE ANALYSIS TO AID ON BREAST DISEASE DIAGNOSIS

Tiago B. Borchartt, tbonini@ic.uff.br
Roger Resmini, rresmini@ic.uff.br
Aura Conci, aconci@ic.uff.br
Federal Fluminense University, R. Voluntário da Pátria, 156, Boa Viagem, Niterói - RJ

Alex Martins, hilexm@gmail.com
Aristófanes C. Silva, ari@dee.ufma.br
Edgar M. Diniz, edgkff@gmail.com
Anselmo Paiva, paiva@dinf.ufma.br
Federal University of Maranhão, Av. Dos Portugueses, São Luíz - MA

Rita C. F. Lima, ritalima@ufpe.br
Federal University of Pernambuco, Av. Prof. Moraes Rego, 1235 - Cidade Universitária, Recife - PE

Abstract. *The human body is a system which looks for thermal symmetry. Non symmetrical temperature distribution in this system can sometimes represent some abnormality or even a disease. Body's skin is the organ responsible for thermoregulation between body core and external environment. Normally this temperature is taken locally (axillary, tympanic, oral, rectal or esophagic) and assumed as global temperature. The thermal infrared camera measures the infrared radiation emitted by the body and converts it into a digital image. Such image is called thermogram which can be used for some analyses and diagnosis aid.*

In this work we apply thermal image analysis to detect difference of temperature between breasts intending to identify breast cancer indicatives. This cancer presents the highest incidence in women according to INCA (National Cancer Institute). Mammography is nowadays accepted by experts as the gold standard for breast cancer detection due to high accuracy for detection of tumors. However, this examination has some negative characteristics like: patient's discomfort during the examination; difficulty of identifying suspicious areas in dense breasts and possible rupture of tumors due to breast compression, releasing cancer cells into the bloodstream. For these reasons, thermography could improve early detection of breast diseases. Thermography does not emits ionizing radiation, is painless and can detect the difference in the breast's thermal pattern up to ten years before the occurrence of cancer.

This work aims to help on diseases diagnosis by thermogram analyses applying a three-step approach. In the first step, thermal images are segmented and separated on left and right breast regions. After segmentation (second step), some features are extracted: range temperature, mean temperature, standard deviation and the quantization of higher tone in an eight level posterization. This last feature considers the entire image temperature and calculates the percentage of area occupied by pixels with the higher temperatures of the image. In the third step, a supervised learning method based on support vector machine (SVM) was used for the extracted feature classification. The features were extracted from a set of 28 images confirmed by physician diagnose. The proposed method achieved the average results of accuracy 85.71%, sensitivity 95.83% and specificity 25.00%.

Keywords*: thermal image, breast diseases, feature extraction, basic statistic measures*

1. INTRODUCTION

A neoplasm originates from cells that suffer some genetic mutation and they begin to reproduce uncontrollably. This neoplasm requests a neo-angiogenesis to nourish its cells. Angiogenesis (Fox et al., 1997) is a phenomenon in which the body creates blood vessels providing an increased flow of blood in the region and therefore an increase in regional temperature (INCA, 2011).

The human body is a system that exchanges heat with the environment (radiation) to ensure a very small variation in temperature in the vital organs such as brain and heart. This process is called thermoregulation. The skin is the organ that mediates the exchange of heat with the external environment. In general, the temperature is measured at specific points on the skin, such as armpits and mouth, and estimated as the overall temperature of the body. But the body has different degrees of heat exchange. Thermic radiation occurs in the infrared range and can be captured through infrared thermal camera which is a device capable of measuring the temperature of various points in a scene (Qi et al., 2001).

To detect the presence of breast neoplasm, this work considers to detect the patient asymmetry. For this it compares each thermal images of the breast, left and right, in the same was physician as the analysis the mammography images (Conci et al., 2010b; Conci et al., 2010c).

This work proposes a new methodology for detecting breast diseases based on the comparison of thermal images of the left and right breast. Such methodology is based on the extraction of simple statistics features: Average intensity, standard deviation, the difference between higher and lower intensity of the gray levels of the image and the percent of the pixels of the last level of the image posterization in eight level. Two approaches are used: the entire image of the breast and the image of the breast divided into four equal parts.

The work is presented in four sections: Related Works, Methodology, Results and Conclusion. The first section shows the most recent works on diagnosis of breast diseases using thermal image; The second is divided into three parts: pre-processing, feature extraction and classification, where we use a classifier that implements a support vector machine (SVM) (Chang and Lin, 2001); In the third section the results are exposed and compared to some other methods; In the last section (Conclusion), the results are evaluated and future work are proposed.

2. RELATED WORKS

Schaefer et al. (2009) consider that the automatic segmentation methods developed until the date of the work were not sufficiently good to segmentation of thermograms. The paper has no information about the database used. The images were manually segmented by experts, generating two images of interest (ROI), left breast and right breast.

Thirty-eight features were extracted as described to follow:

- Basic statistical features: The absolute difference of the temperature mean, the standard deviation, the median temperature and the 90-percentile of left and right breast (four features);
- Image moments: The absolute difference between the center of gravity, the geometric center, m01 and m10 moments of left and right breasts (four features);
- Normalized histogram of both ROIs: the cross-correlation between the two histograms, from the difference histogram they compute the absolute value of its maximum, the number of bins exceeding a threshold (0.01, empirically chosen), the number of zero crossings, energy and the difference of the positive and negative parts of the histogram (eight features);
- Cross co-occurrence matrix: Homogeneity, energy, contrast, symmetry, the first four moments m1–m4 of the matrix (eight features);
- Mutual information: the sum of the breast left entropy plus the breast right entropy plus the joint entropy (one feature);
- Fourier spectrum: the maximum difference of the values of the spectrum and the maximum distance this position of spectrum to the center of the graphic (three features);
- After they apply a Laplacian filter to enhance image contrast and extract again the features of the resulting image: the Fourier features, mutual information and the eight features of cross co-occurrence matrix (twelve features).

The last step describes the pattern classification from the features extracted using fuzzy logic. The best result was with fourteen partitions where obtained accuracy was 79.53%, sensitivity was 79.86% and specificity was 79.49%.

Serrano et al. (2010) describes a methodology of feature extraction based on fractal geometry: Hurst coefficient and Lacunarity. This work is organized in three steps: preprocessing, feature extraction and classification. In the preprocessing step, the work performs a manual segmentation and a resample (100 x 100 pixels). In the feature extraction are extracted thirty-six features using the Hurst coefficient and ninety-seven features using Lacunarity. They arrange the features in fourteen groups. After, the authors utilize seventy-six classifiers, all from the Weka software (Hall et al. 2009). Finally, they evaluate the best result using the analysis of the area under the curve of the Receiver's Operating Characteristics (ROC), Tab. 1 show this, where *G01* is the group containing all features extracted by Serrano et al. (2010), *G02* contains only features extracted using Hurst coefficient and *G09* contains only features extracted using Lacunarity. In this work we used the same image dataset of the Serrano et al. (2010).

Table 1. Serrano et al. (2010) results of the area under the ROC curve.

Techniques	*G01*	*G02*	*G09*
Naive Bayes	0.708	0.875	0.490
Naive Bayes Simple	0.792	0.854	0.557
Naive Bayes Updateable	0.708	0.875	0.490

3. METHODOLOGY

This section presents the proposed methodology for the feature extraction and analysis. Its have three steps: preprocessing, feature extraction and classification.

3.1. Preprocessing

There are various ways to segment images, some using threshold, edge and region based techniques (Conci et al., 2010a). In this work, the method proposed by Motta et al. (2010) is used. This method uses threshold and edge based techniques, it applies Hough transform, Canny edge detection and other techniques.

The images used have 320x240 pixels. Temperatures are initially presented in false-color, as shown in Fig. 1(a), where each color represents a different temperature range. Using the software of the thermal camera the false-color images are transformed to grayscale images, Fig. 1(b).

After converted to grayscale, the image pass through the automatic segmentation algorithm proposed by Motta et al. (2010), resulting in to ROIs the left and right breast, Fig. 1(c) and 1(d). Manually refinement is applied in the ROIs. This refinement aims to delete all contents of segmented image that not belonging to the breast. The result of this preprocessing can be seen in Fig. 1(e).

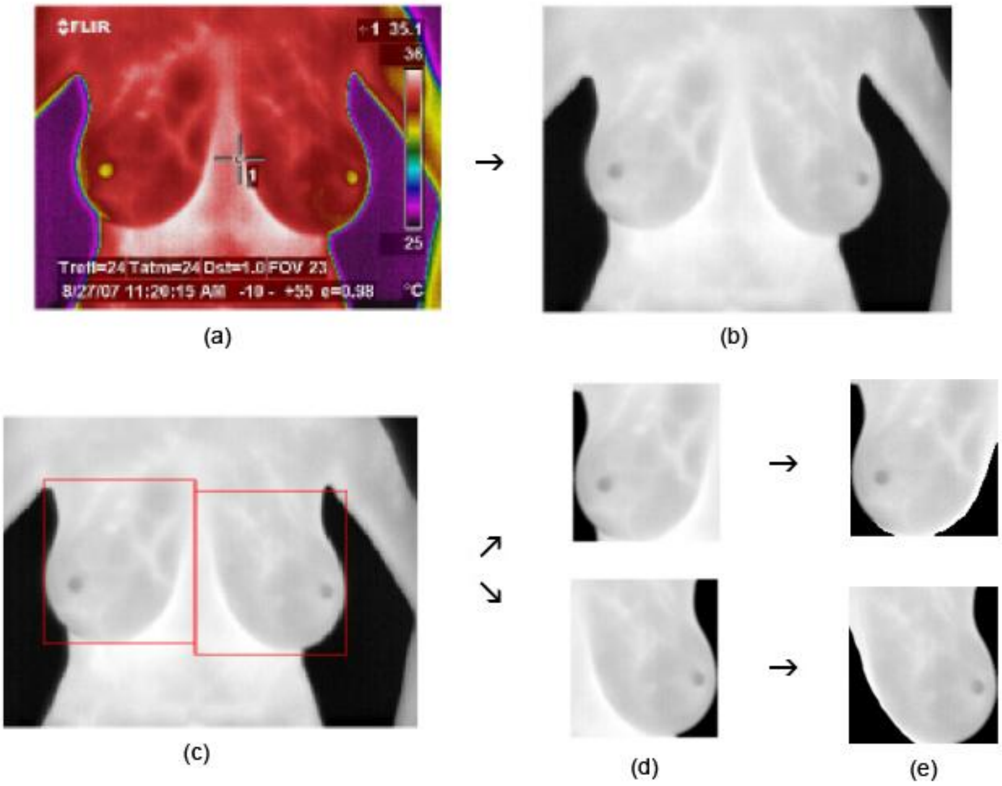
FLIR
1 35.1
36
25
Trefl=24 Tatm=24 Dst=1.0 FOV 23
8/27/07 11:20:15 AM -10 - +55 e=0.98
°C
(a)
(b)
(c)
(d)
(e)

Figure 1. (a) original false-color image; (b) conversion to grayscale; (c) Motta et al. (2010) segmention; (d) separated ROIs; (e) manually refined ROIs.

The automatic segmentation method proposed by Motta et al. (2010) uses as reference the armpits and inframammary fold to perform the segmentation and separation of the breasts. The refinement performed manually after the automatic segmentation uses as reference the internal contour of the mamma, removing the area that is not belonging in the breast.

3.2. Feature Extraction

The features selected for analysis are simple statistics: the range of temperature in the ROI, the mean temperature, the standard deviation and the quantization of the higher tone in an eight level posterization. Two different approaches are used for feature extraction. In the first, the entire image is used. In the second, the ROIs are divided into four. In this division into quadrants, the ROI is divided in four parts with the same size (i.e. the proportion of each quadrant is reduced to one quarter (1/4) of the original rectangle).

The Range measures the difference between the pixels of greater and lesser intensity, excluding the background. This measure indicates the thermal variation within the ROI, it is expected that breast pathology has a temperature range higher than a healthy breast. Equation (1) shows this feature, where p_{ij} is the intensity of the pixel in the position (i, j) in the ROI.

$$Range = \left|Max(p_{ij}) - Min(p_{ij})\right| \tag{1}$$

The mean is the statistic that indicates the intensity more frequency in the image. The difference of the means of the patients ROIs with disease is higher than a healthy patient. Equation (2) shows the formula used to calculate the mean.

$$\mu = \frac{1}{N.M}\sum_{j=1}^{N}\sum_{i=1}^{M} p_{ij} \tag{2}$$

The standard deviation shows the dispersion of the ROI pixels intensity from the mean. A low standard deviation indicates that the data points tend to be very close to the mean, whereas high standard deviation indicates that the data are spread out over a large range of values. Equation (3) shows the formula used to calculate the standard deviation.

$$S = \sqrt{\frac{1}{N.M}\sum_{j=1}^{N}\sum_{i=1}^{M}\left(p_{ij} - \bar{\mu}\right)^2} \tag{3}$$

Quantization of the higher level in an eight level posterization: to compute this feature, firstly the histogram of each ROI is calculated. Then a posterization in eight levels is made (this number of levels is chosen by testing). We verify that the area of the last level represents the regions with greater intensities of gray level of the ROI, i. e., with higher temperatures. When compared a breast healthy and a pathological, it is observed this feature is evidently greater where there is pathology. Figure 2 shows the steps of the process.

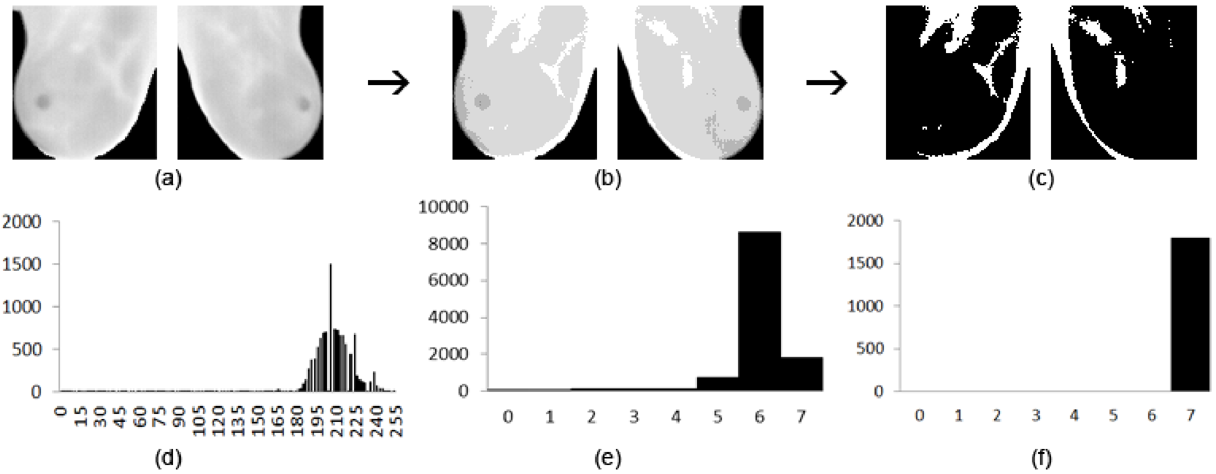

Figure 2. (a) Segmented ROI; (b) Posterization in eight levels; (c) Quantization of the higher level of the posterization; (d) Histogram of image (a); (e) Histogram of 8-levels image (b); (f) Histogram of image (c)

3.3. Classification

We made the classification of features to aid in the diagnosis of patients from the thermal images through the free software LibSVM. SVM is a method of supervised machine learning, known as binary linear classifier not probabilistic. The decision by the SVM is based on the construction of a hyperplane or a set of hyperplanes in a high dimensional space or infinity. This n-dimensional space can be used for regression, classification, or other tasks. The objective of this type of classification is to devise a computationally efficient way to maximize the margins between the data. Thus it is expected to improve the generalization of the data set analyzed (Nunes et al., 2010).

In practice, given a training dataset *X* with data from two classes, SVM separates these classes in a set of hyperplanes determined by the data of *X*, called support vectors. These hyperplanes maximizes the margin, or increase the distance of each class. A good separation of data is obtained by the hyperplane which has the largest distance between classes (the margin functional). Figure 3 shows an example of two classes (represented by circles and triangles) separated by a margin functional (two solid lines) separating the two hyperplanes (dotted line).

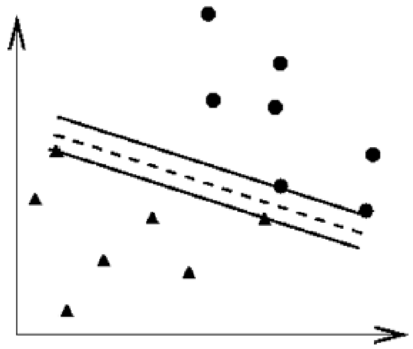

Figure 3. A SVM typical classification in two classes

We use a support vector classification for non-uniform data (NU-SVC) and a radial kernel (radial basis function) chosen from tests. In this work we use two approaches to the extraction and analysis of the features: using the entire image; and using the image divided into quadrants.

In the first approach we use the entire image for feature extraction. To perform the analysis of these data, we made the difference between the features in the left and right breast to estimate the contralateral symmetry. When the images are relatively symmetrical, small asymmetries may indicate a suspicious region.

In the second approach we use the image partitioned into quadrants to extract features. The image was partitioned for data analysis to be more sensitive to small local differences. Due the image resolution, four quadrants are appropriates because they ensure that each quadrant contain a relevant information. In this approach, the extracted features are analyzed separately for each breast.

4. RESULTS

After feature extraction, these data were organized as the file-standard input format for LibSVM software. The SVM classified the images in two classes (with pathology and healthy). We analyzed twenty-eight segmented images as described in Section 3.1. Four of these images are without pathology and twenty-four present diagnosed with the disease.

We used the method of leave-one-out to analyze the features in both approaches. In the algorithm of leave-one-out, *N-1* instances are used to train the model and this is validated by testing it on the instance left out. The experiment is repeated for a total of N times, each time leaving out a different instance for validation (Burnham, 2004). Fundamentally, the average specificity and sensitivity for all N interactions are used to evaluate the performance of the SVM model used in this work.

To calculate accuracy, sensitivity and specificity we use the data in the confusion matrix, which is the amount of True-Positive (TP) True-Negative (TN), False Positive (FP) and False Negative (FN), where:

- True-Positive: Disease patients correctly classified as diseased;
- True-Negative: Healthy patients correctly identified as healthy;
- False-Positive: Healthy patients incorrectly classified as diseased;
- False-Negative: Disease patients incorrectly identified as healthy;
- Sensitivity: measures the proportion of positive cases which are correctly identified as positive, as shown in Eq. (4);
- Specificity: measures the proportion of negative cases which are correctly identified as negative, as shown in Eq. (5);
- Accuracy: percentage of correct classification, as shown in Eq. (6).

The calculation of the measures is done as follows:

$$Sensitivity = \frac{TP}{TP + FN} \tag{4}$$

$$Specificity = \frac{TN}{TN + FP} \tag{5}$$

$$Accuracy = \frac{TP + FP}{TP + FP + TN + FN} \tag{6}$$

In the first approach (that uses the entire image to extract features), we extract four features for each patient (range, mean, standard deviation and quantization of the eight level) and we calculate the feature value difference of the left and right breast, organizing a feature vector for each patient, as can be seen in Tab. 2. All the elements of this vector are normalized between 0 and 1.

Table 2. Extracted features distribution from the entire breast image (left breast minus right breast).

Feature Patient ID	Diagnosis	Range	Mean	Standard Deviation	Quantization
IR_0086	Pathological	0	2.690	1.274	0.092
IR_0096	Pathological	56	5.153	6.919	0.313
IR_0146	Healthy	38	11.957	1.069	0.116
IR_0713	Pathological	70	6.640	10.381	0.439
IR_0753	Pathological	70	2.367	10.733	0.087
IR_1020	Pathological	33	12.446	3.662	0.070
IR_1027	Pathological	0	19.511	3.691	0.158
IR_1038	Pathological	14	21.104	5.395	0.162
IR_2886	Pathological	1	14.720	3.048	0.065
IR_3434	Pathological	44	20.934	0.878	0.097
IR_3438	Pathological	0	2.1758	0.963	0.625
IR_3600	Pathological	23	19.797	10.321	0.029
IR_3724	Pathological	0	25.368	4.697	0.041
IR_3774	Healthy	0	5.7554	1.044	0.038
IR_3835	Pathological	0	8.101	2.158	0.118
IR_3840	Healthy	0	17.215	3.489	0.140
IR_3924	Pathological	0	2.644	1.802	0.104
IR_4003	Pathological	0	15.675	0.873	0.130
IR_4872	Pathological	0	4.313	2.081	0.717
IR_5314	Pathological	0	4.580	2.736	0.072
IR_5353	Pathological	0	2.688	1.797	0.180
IR_5528	Healthy	57	2.455	5.902	0.082
IR_5560	Pathological	0	2.044	0.865	0.607
IR_5667	Pathological	0	0.847	0.895	0.132
IR_5752	Pathological	0	1.505	0.410	0.115
IR_5908	Pathological	0	2.593	1.386	0.134
IR_5926	Pathological	0	4.668	3.843	0.182
IR_7460	Pathological	0	13.683	2.4319	0.632
IR_0086	Pathological	0	2.690	1.273	0.092
IR_0096	Pathological	56	5.153	6.919	0.313

As can be seen in Tab. 2, the difference values of the quantization of the last tone of a posterization of eight tones allow a separation between the breast with or without pathology. The other extracted features when analyzed separately will not allow this separation.

The results are calculated using a leave-one-out method. The results obtained by using SVM in this first approach can be seen in Tab. 3.

Table 3. Confusion Matrix from the first approach

23 TP	3 FP
1 FN	1 TN

For the second approach we partitioned each segmented image into four quadrants for feature extraction. We extract four features from each quadrant. A total of sixteen features by breast were obtained, i.e., thirty-two per patient. These features were analyzed by SVM, the results can be seen in Tab. 4.

Table 4. Confusion Matrix from the second approach

16 TP	8 FP
3 FN	1 TN

Table 5 show a comparison between the results obtained using both approach and other methods.

Table 5. Results obtained by SVM for the first approach.

	Sensitivity (%)	Specificity (%)	Accuracy (%)
First approach results	95.83	25.00	85.71
Second approach results	66.70	25.00	60.70
Schaefer et al. (2009)	79.86	79.49	79.53

Shown in Tab. 5, the high value of rate sensitivity of the first approach demonstrates that the method achieves a good ranking among the diagnoses with pathology. The low specificity value of the two approaches we believe to be related to unbalanced sample, since it has only four images of healthy patients. The accuracy rate shows that the proposed method correctly classified the diagnosis in most cases.

Comparing the results of the best approach this work with the results obtained by Schaefer et al. (2009) we note that the accuracy and sensitivity of this work were higher. The specificity obtained for this work was inferior we believe that because the sample is unbalanced. Although we know that the databases used and the features analyzed are different.

Table 6 shows the results obtained in this work and results presented by Serrano et al. (2010) analyzing the area under the ROC curve. We note that the results obtained by the first approach presented in this work were better than the results obtained by them when using only Lacunarity. However, it remains lower compared with the features obtained using the Hurst coefficient or associating Lacunarity and Hurst coefficient. The images set utilized in both works are the same.

Table 6. Results obtained by analyzing the area under the ROC curve.

Technique	**Area under ROC curve**
Serrano (2010) Lacunarity and Hurst coefficient	0.708
Serrano (2010) Hurst coefficient	0.875
Serrano (2010) Lacunarity	0.490
First approach in this work (entire image)	0.604
Second approach in this work (divided image in quadrants)	0.458

5. CONCLUSION

The results obtained show the viability of using simple statistical measures as a first approach to aid diagnosis of breast disease by thermal images. The present results show that for the analysis of breast thermography using the entire image for feature extraction are more appropriate than the use of subdivided images. Although the measures used are simple, the methodology shows adequate results.

The method using the entire image when compared with other related works presents superiors values of accuracy and sensitivity but lower specificity. Another comparison is made using analysis of area under the ROC curve, concern this, results from this work are better for Lacunarity measures, but lower when the Hurst coefficient is used.

As future work we suggest the use of other measures, such as texture descriptors, fractal measures (Higuchi, Box Counting) and analyze the frequency domain using Fourier or Wavelet. Another improvement to the methodology would be automating the process of segmentation and refinement of regions of interest in the images.

6. ACKNOWLEDGEMENTS

This research is supported by Coordination of Improvement of Higher Education Personnel - Brazilian Education Ministry, Program Pro-Engineering PE no. 021-2008 (in Portuguese: CAPES, Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Ministério da Educação, Programa Pró-Engenharias PE021-2008).

This research is also partially financed by National Council for Scientific and Technological Development (in Portuguese: CNPq, Conselho Nacional de Desenvolvimento Científico e Tecnológico.)

7. REFERENCES

Burnham, K.P., Anderson D., 2004. "Model Selection and Multi-Model Inference." Berlim: Springer; 2004.

Chang, C.C. and Lin, C.J., 2001. "LIBSVM: a library for support vector machines." Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>

Conci, A., Lima, R.C.F., Fontes, C.A.P., Motta, L.S., Resmini, R., 2010a. "A new method for automatic segmentation of the region of interest of thermographic breast image." Thermology International, Vol. 20, No. 4, pp 134-135.

Conci, A., Lima, R.C.F., Fontes, C.A.P., Vasconcelos, S., Borchartt, T.B., Resmini, R., 2010b. "On the breast reconstruction by thermal images" Thermology International, Vol. 20, No. 4, 135 p.

Conci, A., Lima, R.C.F., Fontes, C.A.P., Borchartt, T.B., Resmini, R., 2010c. "A new method to aid to the breast diagnosis using fractal geometry." Thermology International, Vol. 20, No. 4, pp 135-136.

Fox, S.B., Leek, R.D., Bliss, J., Mansi, J.L., Gusterson, B., Gatter, K.C., Harris, A.L., 1997. "Association of tumor angiogenese with bone marrow micrometastases in breast cancer patients." Journal of the National Cancer Institute, Vol. 89, No. 14, pp 1044-1049.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H., (2009). "The WEKA Data Mining Software: An Update". SIGKDD Explorations, Vol. 11, Issue 1.

INCA, Instituto Nacional do Câncer, 01 Mar. 2011 <http://www2.inca.gov.br>

Motta, L.S., Conci, A., Lima, R.C.F., Diniz, E.M., 2010. "Automatic segmentation on thermograms in order to aid diagnosis and 2D modeling." Proceedings of 10º Workshop em Informática Médica. Vol. 1, pp 1610-1619.

Ng, E., 2008. "A review of thermography as promising non-invasive detection modality for breast tumor." International Journal of Thermal Sciences, Vol. 48, pp. 849-859.

Nunes, A.P., Silva, A.C., Paiva, A.C., 2010. "Detection of masses in mammographic images using geometry, Simpson's Diversity Index and SVM." International Journal of Signal and Imaging Systems Engineering, Vol. 3, No. 1, pp. 40-51.

Qi, H., Head, J., 2001. "Asymmetry analysis using automatic segmentation and classification for breast cancer detection in thermograms." Proceedings of the 23rd IEEE Annual International Conference on Engineering in Medicine and Biology, vol. 3, pp. 2866-2869.

Schaefer, G., Zavisek, M., Nakashima, T., 2009. "Thermography based breast cancer analysis using statistical features and fuzzy classification." Pattern Recognition Vol. 42, No. 6, pp. 1133-1137.

Serrano, R.C., Ulysses, J., Ribeiro, S., Conci, A., Lima, R.C.F., 2010. "Using Hurst coeficiente and Lacunarity to diagnosis early breast diseases." Proceedings of 17º International Conference on Systems, Signal and Image Processing, Vol. 1, Rio de Janeiro, Brazil, pp. 550-553.

8. RESPONSIBILITY NOTICE

The authors are the only responsible for the printed material included in this paper.