Meaningful Features for Computerized Detection of Breast Cancer

José Anibal Arias, Verónica Rodríguez, and Rosebet Miranda

Universidad Tecnológica de la Mixteca, Km 2.5 Carretera a Acatlima
CP 69000 Huajuapan de León, Oaxaca, México
{anibal,veromix,rmiranda}@mixteco.utm.mx

Abstract. After pre-processing and segmenting suspicious masses in mammographies based on the Top-Hat and Markov Random Fields methods, we developed a mass-detection algorithm that uses gray level co-occurrence matrices, gray level difference statistics, gray level run length statistics, shape descriptors and intensity parameters as the entry of a vector support machine classifier. During the classification process we test up to 63 image features, keeping the 35 most important and obtaining 85% of accuracy score.

Keywords: Breast cancer, CADx, Image features, SVM.

1 Introduction

Breast cancer is a disease in which malignant cells grow in breast tissue. This type of cancer is more frequent in middle age women (40-49 years-old) [1] and, in Mexico, it is the primary cause of death from malignant tumors among women [7]. Mammography (X-ray picture of the breast) associated with clinical breast examination is the cheapest and most efficient method for early detection of breast cancer. Radiologists make a visual examination of mammographies searching for masses, calcifications, density asymmetries and structure distortions that reveal the presence of cancer. However, it is very difficult to search for abnormalities because of the small differences in the image densities of breast tissue and the vast range of possible abnormalities, so the task remains highly subjective and qualitative, depending mainly on the quality of the mammography and the training and experience of radiologists [10]. This is a risk, especially in third level developed countries, where there are no other diagnosis protocols widely available.

Computer-aided diagnosis (CADx) is a helpful tool that improves diagnostic accuracy assisting radiologists to make correct mammography interpretation. The detection sensitivity without CADx is around 80% and with it up to 90% [6]. The tasks a CADx system should accomplish are:

Pre-processing. Noise in the digitized mammogram is reduced and the general image quality is improved. Labels, tape and scanning artefacts, and pectoral muscle are removed.

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part II, LNCS 8259, pp. 198–205, 2013.

Segmentation. Suspicious regions are isolated to be later classified as abnormality (true positive) or tissue (false positive).

Feature Extraction. Several features are obtained from the suspicious regions.

Classification. CADx system declares each detected region as an abnormality or normal breast tissue. Also, in this stage, if the region is an abnormality, their malignant or benign class is determined.

Several CADx systems have been developed for research purposes [15], but there is no report of any commercial system available. We intend to develop one for detection and diagnosis of masses (in a first version, identifying other abnormalities later) and make it available to public health institutions. In this work, we present the last two stages of a CADx system that identifies masses in mammographies. Masses are subtle areas (2-30 mm in diameter) with smooth boundaries and high densities and represent the most difficult type of lesion to detect and characterize.

The paper is organized as follows. In Section 2, we describe several approaches for automated detection and classification of masses in mammograms. The data used in our tests is mentioned in Section 3. The different features generated from suspicious regions are described in Section 4. The classifier and the experimental results are presented in Section 5. Finally, conclusions and future work are given in Section 6.

2 Related Work

Several methods have been proposed for mammography mass detection. Excellent state of art reviews are presented in [11] and [2], showing an evaluation of several methods for enhancement of mammographic images, detection and classification of masses.

Rojas and Nandi [13] proposed a three stages method to perform mass detection. The first one is a multilevel adaptative process based on local statistical measure of the pixel intensities and morphological operators to enhance breast structures. In the next stage, the images are segmented by applying thresholding and Gaussian filtering. Finally, the selection of suspicious regions is performed by means of a ranking system that uses 18 shape and intensity features. The method was tested on 57 mammographic images of masses from the MIAS database [17], and achieved a sensitivity of 80% at 2.3 false-positives per image.

An interesting method for reduction of false positives in mass detection is presented by Llado et al. [9]. The basic idea of their approach is the use of Local Binary Patterns for texture descriptions of ROIs. Support Vector Machines (SVM) with a polynomial kernel performed classification of mass and normal breast tissue. Their approach was evaluated on 1792 ROIs extracted from the DDSM mamographic database [5], and reported a mean Az value (area under the ROC curve) of 0.94.

Sampaio et al. [14] proposed a methodology based on Cellular Neural Networks, geostatistic functions and Support Vector Machines. In the first step

of their methodology, the images are pre-processed by using Hough Transform, K-means and morphological operators. Identification of suspicious regions is performed by segmentation with Cellular Neural Networks. A SVM classifier that uses shape and texture features is proposed with a sensitivity of 80% at 0.84 false positives per image.

3 Database

Our method was tested on a subset of images extracted from the Mammographic Image Analysis Society (MIAS) database [17]. This publicly available digitized database contains left and right breast images in mediolateral oblique (MLO) view that represent the mammograms of 161 patients with ages between 50 and 65. All images were digitized at a resolution of 1024×1024 pixels and at 8-bit gray scale level.

The chosen set corresponds to masses annotated as spiculated, circumscribed or miscellaneous (ill-defined masses). The summary of this dataset by type of mass and density of breast tissue is shown in Table 1.

Table 1. Summary of MIAS images used

	Fatty	Fatty-Glandular	Dense-Glandular	Total
Circumscribed	13	8	3	24
Miscellaneous	8	5	2	15
Spiculated	5	7	7	19
Total				58

For decreasing computational cost, all images were reduced by a factor of two. Moreover, the 3×3 median filter was applied to reduce noise, and labels and pectoral muscle were manually extracted from the images with help of the ImageJ program [12]. With the purpose to filter and enhance the contrast of the possible mass regions, the Top-Hat transform was applied to all images. A disk was used as structural element to filter suspicious regions. The size of the disk was iteratively modified from two pixels to the width of breast area. Then, detection of suspicious regions (ROIs) was done by applying segmentation based on texture and Markov Random Fields. A Gaussian observation model with three texture features of first order: mean, standard deviation, and entropy, was used. In total, 278 ROIs of different sizes were identified, from which, 50 represent suspicious masses, while the other 228, normal tissue. These ROIs are the entry to the classification stage.

4 Features

The following stage of mass detection by CADx systems is the feature extraction and selection. The feature space is very large and complex, but only some of features are significant. After years of intensive research, hundreds of features have

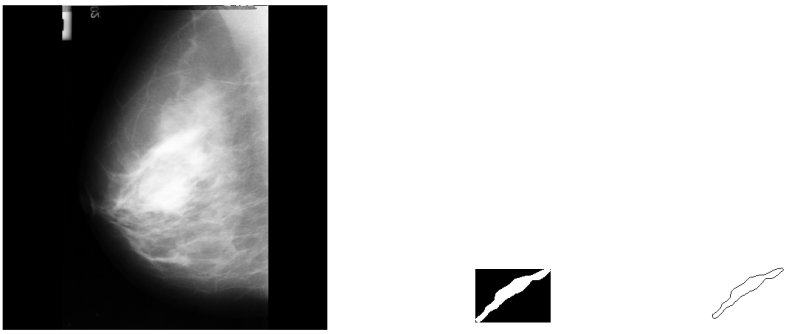
been proposed. But using many features degrades the performance of the classifiers, so that redundant features should be removed to improve the performance of the classifier. There are basically three types of features: intensity, geometric and texture features. After reviewing many feature evaluation initiatives [8], [18], [20], we chose an important and discriminative subset of 35 features for mass detection.

4.1 Intensity Features

Three basic statistics of the detected ROIs were used: skewness, kurtosis and entropy.

4.2 Shape Features

Before the extraction of these features, the detected ROIs are binarized and processed to identify their boundaries. In Fig. 1 some examples of results for these processes are shown. Seven features were directly calculated from the pixels in the boundaries and within area of ROIs: perimeter, area, compactness, and the first four central invariant moments.



(a) Original MIAS image (b) Binarized ROI (c) ROI shape

Fig. 1. ROI processing for shape features extraction

4.3 Texture Features

Texture is the term used to characterize the surface of a given region, and it is one of the main features used in identifying ROIs in an image [3]. In general, texture features can be grouped into three classes based on what they are derived from: Gray-level co-occurrence matrices, Gray-level difference statistics, and Gray-level run length statistics.

Gray-Level Co-occurrence Matrix (GLCM). An element of the GLCM matrix $P(i, j, d, \theta)$ is defined as the joint probability that the gray levels i and j occur separated by a distance d and along direction θ of the image [2]. Four GLCM matrices were calculated from each ROI using $\theta = \{0°, 45°, 90°, 135°\}$ and $d = 1$. From these matrices, the six following features were obtained (and averaged in the four directions): contrast, correlation, variance, energy, entropy and homogeneity.

Gray-Level Difference Statistics (GLDS). The GLDS vector is the histogram of the absolute difference of pixel pairs which are separated by a given displacement δ [19]. Also, to obtain GLDS features, four forms of the vector δ were considered: $(0, d), (-d, d), (d, 0)$, and $(-d, -d)$. Three textural features were measured and averaged (considering $d = 1$) from these vectors: mean, entropy and variance.

Gray-Level Run Length Statistics (GRLS). The GRLS method is based on computing the number of gray-level runs of various lengths [4]. A gray-level run is a set of consecutive and collinear pixel points having the same gray-level value. The length of the run is the number of pixels in the run. For an $M \times N$ run length matrix $p(i, j)$, M is the number of gray levels and N is the maximum run length. In a study [4], four feature extraction functions following the idea of joint statistical measure of gray level and run length gave better performance:

1. Short run low gray level emphasis (SRLGE)

$$SRLGE = \frac{1}{n_r} \sum_{i=1}^{M} \sum_{j=1}^{N} \frac{p(i,j)}{i^2 \cdot j^2} \tag{1}$$

2. Short run high gray level emphasis (SRHGE)

$$SRHGE = \frac{1}{n_r} \sum_{i=1}^{M} \sum_{j=1}^{N} \frac{p(i,j) \cdot i^2}{j^2} \tag{2}$$

3. Long run low gray level emphasis (LRLGE)

$$LRLGE = \frac{1}{n_r} \sum_{i=1}^{M} \sum_{j=1}^{N} \frac{p(i,j) \cdot j^2}{i^2} \tag{3}$$

4. Long run high gray level emphasis (LRHGE)

$$LRHGE = \frac{1}{n_r} \sum_{i=1}^{M} \sum_{j=1}^{N} p(i,j) \cdot i^2 \cdot j^2 \tag{4}$$

where n_r is the total number of runs.

These four features were calculated in four positive directions: $0°, 45°, 90°$ and $135°$ (16 features) for Test 2. For Test 1 we add seven more features calculated in four directions (28 features): Short Run Emphasis (SRE), Long Run Emphasis (LRE), Gray-Level Nonuniformity (GLN), Run Length Nonuniformity (RLN), Run Percentage (RP), Low Gray-Level Run Emphasis (LGRE), High Gray-Level Run Emphasis (HGRE).

5 Classification and Experimental Results

5.1 Support Vector Machine (SVM)

SVM classifier [16] is a relative new option for doing classification. It has their roots in the existence of an optimal (in the sense of quadratic convex optimization) hyperplane that separates two classes. Data is projected by means of a kernel function in a high-dimensional space and, in this space, the hyperplane is linear, but their projection back in original space is non-linear. In our experiments we use a radial basis function (RBF) kernel. The fit of the hyperplane to data is controlled by the parameter β of the RBF function and the SVM parameter C that controls the width of the classifier's margin.

5.2 Experiments

For the experiments, the set of 278 detected ROIs was randomly divided in 25 masses and 114 normal tissue segments for training, and the equivalent for testing. In the first experiment (Test 1) we tested the 63 intensity, shape and texture features described in Section 4; the corresponding results are presented in Table 2. In other experiments we tested different subsets of texture features, and the best results were obtained with the first 35 features mentioned in Section 4.

Table 2. SVM classification results using a RBF kernel and the full set of 63 features

Parameters	Accuracy in training set	Number of support vectors	Accuracy in test set
$\beta = 1, C = 2$	92.19 %	58	**84.18 %**
$\beta = 1, C = 10$	98.57 %	62	79.86 %
$\beta = 2, C = 2$	76.98 %	16	71.95 %
$\beta = 0.5, C = 2$	98.57 %	98	84.18 %
$\beta = 1, C = 1$	50.4 %	10	52.52 %

SVM classifier gave the best results with $\beta = 1$ and $C = 2$; defining 54 support vectors (Table 3). Differences in the results represent the compromise between accuracy in test stage and number of support vectors. We tested different subsets of texture parameters and different kernels, but we are reporting here the best scores.

Table 3. SVM classification results using a RBF kernel and the best 35 features

Parameters	Accuracy in training set	Number of support vectors	Accuracy in test set
$\beta = 1, C = 2$	88.49 %	54	**84.18 %**
$\beta = 1, C = 10$	94.25 %	56	78.5 %
$\beta = 2, C = 2$	49.7 %	8	48.3 %
$\beta = 0.5, C = 2$	94.25 %	80	84.9 %
$\beta = 1, C = 1$	74.2 %	12	71.3 %

6 Conclusions and Future Work

We selected and tested some of the simplest and most discriminant features for digital processing of mammographies. After pre-processing and segmenting the ROIs of the MIAS database, SVM classification gives reasonably good accuracy scores with only 35 well known features.

With this framework we can test more shape and texture features, as well as classifiers and combinations among them. We are still far away of our purpuse, but with the future improvement of the different stages, we will be closer to build a working CADx system available for public service.

Acknowledgments. This research is supported by the mexican SEP-PROMEP program within the project "Detección, Segmentación e Identificación de Masas en Imágenes de Mamografía". 103.5/12/4621, (IDCA: 10151 CLAVE: UTMIX-CA-32).

References

1. Brandan, M., Villaseñor, N.: Detección del cáncer de mama: Estado de la mamografía en México. Cancerología: Revista del Instituto Nacional de Cancerología de México 1(3), 147–162 (2006)
2. Cheng, H., Shi, X., Min, R., Hu, L., Cai, X., Du, H.: Approaches for automated detection and classification of masses in mammograms. Pattern Recognition 20, 646–668 (2006)
3. Chang, T., Kuo, C.: Texture Analysis and classification with tree-structured wavelet transform. IEEE Transactions on Image Processing 2, 429–441 (1993)
4. Dasarathy, B., Holder, E.: Image characterizations based on joint gray-level run-length distributions. Pattern Recognition Letters 12, 497–502 (1991)
5. Heath, M., Bowyer, K., Kopans, D., Moore, R., Kegelmeyer, W.P.: The digital database for screening mammography. In: Proceedings of the Fifth International Workshop on Digital Mammography, pp. 212–218. Medical Physics Publishing (2001)
6. Horsch, A., Hapfelmeier, A.: Needs assessment for next generation computer-aided mammography reference image databases and evaluation studies. International Journal of Computer Assisted Radiology and Surgery 6(6), 749–767 (2011)

7. Instituto Nacional de Estadística, Geografía e Informática (INEGI): Estadística a propósito del día mundial contra el cáncer. Boletín de Prensa (Febrero 2010), http://www.inegi.org.mx/inegi/contenidos/espanol/prensa/Contenidos/estadisticas/2010/cancer10.doc
8. Kim, J., Park, H.: Statistical Textual Features for Detection of Microcalcifications in Digitized Mammograms. IEEE Transactions on Medical Imaging 18(3), 231–238 (1999)
9. Lladó, X., Olivier, A.: A textural approach for mass false positive reduction in mammography. Computerized Medical Imaging and Graphics 33(6), 415–422 (2009)
10. Li, H., Liu, Y., Lo, S., Freedman, M.: Computerized Radiographic Mass Detection-Part II: Decision Support by Featured Database Visualization and Modular Neural Networks. IEEE Transactions on Medical Imaging 20(4), 302–313 (2001)
11. Oliver, A., Freixenet, J.: A review of automatic mass detection and segmentation in mammographic images. Medical Image Analysis 14(2), 87–110 (2010)
12. Rasband, W.: ImageJ. National Intitutes of Health, USA (1997), http://imagej.nih.gov/ij
13. Rojas, D., Nandi, K.: Detection of masses in mammograms via statistically based enhancement, multilevel-thresholding segmentation, and region selection. Computerized Medical Imaging and Graphics 32(4), 304–315 (2008)
14. Sampaio, B., Moraes, D.: Detection of masses in mammogram images using CNN, geostatistic functions and SVM. Computers in Biology and Medicine 41(8), 653–664 (2011)
15. Samulski, M., Karssemeijer, N.: Optimizing case-based detection performance in a multiview CAD system for mammography. IEEE Transactions on Medical Imaging 30(4), 1001–1009 (2011)
16. Scholkopf, B., Smola, A.: Learning with Kernels. The MIT Press (2002)
17. Suckling, J.: The Mammographic Image Analysis Society Digital Mammogram Database. Exerpta Medica. International Congress Series 1069, 375–378 (1994)
18. Tang, X.: Texture Information in Run-Length Matrices. IEEE Transactions on Image Processing 7(11), 1602–1609 (1998)
19. Weszka, J., Dyer, C., Rosenfeld, A.: A comparative study of texture measures for terrain classification. IEEE Transactions Syst., Man, Cybern. SMC-6, 269–285 (1976)
20. Yin, F., Giger, M., Doi, K., Vyborny, C., Schmidt, R.: Computerized detection of masses in digital mammograms: investigation of feature-analysis techniques. J. Digital Imaging 7, 18–26 (1994)