# **Tumor Classification on Mammographies Based on BPNN and Sobel Filter**

Enrique Calot, Hernan Merlino, Claudio Rancan, and Ramon Garcia-Martinez**<sup>1</sup>**

**Abstract.** Breast cancer is a very common disease and the cause of death of many people. It has been proven that prevention decreases the death rate, but the costs of diagnosis and image processing are very high when applied to all the population with potential risk. This paper studies an existent computer aided diagnosis method using neural network and improves its detection success rate from 60% to 73%. This improvement is achieved due to the use of image and statistical operators over concentric regions around the tumor boundaries.

# **1 State of the Art**

-

Breast cancer is the most common cancer among women worldwide representing a 31% of all tumors in the female population [1, 11]. Screening by mammography alone, with or without physical examination of the breasts, plus follow-up of individuals with positive or suspicious findings, will reduce mortality from breast cancer by up to one-third among women aged from 50 to 69 years. Unfortunately, mammography is an expensive test that requires great care and expertise both to perform and in the interpretation of results. It is therefore currently not a viable option for many countries [8, 9, 11]. Computer systems set on the classification of tumors found on mammographies may be of greater help during the diagnosis process in general hospitals [2, 5].

# **2 Addressed Problem in This Paper**

Tumor classification issue has been addressed using neural networks in a previous work by [5] with a success rate of 60%. In the present work some complementary workarounds had been done to improve that results.

Hernan Merlino, Claudio Rancan, and Ramon Garcia-Martinez University of La Plata, Computer Science School, PhD Program Calles 50 y 120, (B1900) La Plata, Buenos Aires, Argentina e-mail: hmerlino@gmail.com, crancan@itba.edu.ar

Enrique Calot, Hernan Merlino, and Ramon Garcia-Martinez University of Buenos Aires, School of Engineering, Intelligent Systems Laboratory Paseo Colon 850, (C1063ACV) Buenos Aires, Argentina e-mail:{ecalot,rgarciamar}@fi.uba.ar

#### **3 Proposed Workarounds**

The first workaround was the use of *The Digital Database for Screening Mammography* (DDSM), a compilation made by the *University of South Florida* (USF) as a test-and-train database [6, 7].

Improvements available in this database

- Provides images with a higher resolution.
- Contains a bigger number of images.
- Has additional information at disposal. *i.e.* tumor boundaries.

The second workaround is the use of an image filter allowing gradient changes to perform efficient boundary detection.

The third workaround is the analysis of the anomaly as is, and not the whole breast. In order to do that the additional tumor boundary information provided by the database must be used. In the production stage, when new acquired images are to be classified, a method to automatically shape the tumors is to be used. The implementation of such algorithm is not in the scope of this paper.

#### *3.1 Generating Boundary Regions*

A region is defined as a section of the image surface with a specified distance to the boundary. Having known the shape of the anomaly, it is possible to generate a set of concentric regions and another one for the core of the tumor. This work has used 10 regions inside and outside the boundaries and another one in the middle as the tumor core. The usage of regions is very important because the boundary gradient is normal to the region width and providing a good set up for the Sobel operator explained below.

Having a list of pixels belonging to the boundary there are many possible algorithms to calculate adjacent regions. One possible way is applying the Bellman-Ford [3] or the even faster Dijkstra [4] algorithms considering the shape boundary bitmap as a graph where each pixel has a connection weighted 1 to the four immediately adjacent ones and the boundary pixels as start points with null distance between them. Since both weights are non-negative numbers, the Dijkstra algorithm should be preferred. This way to separate regions is very fast and may be stopped when the desired number of regions is reached. There is no need to process all the pixels. The most important disadvantages of this algorithm are that it uses the taxicab 1-norm metric as distance measure and the boundary curve must be a closed shape.

Another way to achieve region separation is applying a Gaussian filter to the shape bitmap. With the resultant image it is possible to define brightness intervals with respective regions associated to them. The Gaussian radius should be considerably as big as to generate distant regions. This algorithm is using a Euclidean

2-norm measure of distance but requires more time to calculate the convolution between the big-enough Gaussian matrix and the whole input image.

#### *3.2 Sobel Operator*

Sobel operator is an image filter capable of transforming an image into two others representing the increment of brightness in each pixel compared to the surrounding ones. Both generated images may be represented either in Cartesian coordinates  $(x, y)$  or in polar ones (module, argument). As a result, the Sobel module image will show brighter pixels on abrupt changes and the Sobel argument image will show the direction of this change [10].

Malignant tumors (cancer) spread themselves to other placements (metastases) and to adjacent tissues (local invasion) causing potential danger. In the other hand, benign tumors do not spread to other locations, having more defined boundaries. The fact that cancer invades adjacent tissues makes their boundaries not so abrupt in images and therefore much more sensitive to a Sobel module filter.

The proposed workaround calculates mean bright, variance and size for a specified region and image. In figure 1, cases *a*, *b*, *c* and *d* have the same bright mean. Furthermore, cases *b* and *c* have the same variance, bigger than *a*; this shows that the variance may be important. Even though, the case *b* and *c* have the same variance, the Sobel module mean is bigger in case *c*, because there are more abrupt changes. Exaggerating an example of a tumor, *b* should correspond to a benign tumor, with a defined boundary and *c* or *d* to a malignant one, with larger borders in a same-sized surface or branches spreading to adjacent tissues.

Some final generated images are shown as figures 2 to 4, where the original, Sobel argument and Sobel module are shown respectively. Note that the original image is not directly the one obtained from the database, it has been pre-processed to drop out the radiographic labels.

After images are generated, they are separated in different regions according to the shape of the tumor and then the mean, variance and size of each resultant region is calculated. The input of the back propagation neural network should be the real numbers obtained in this process.



**Fig. 1** Bright distribution over a region



**Fig. 2** Original image **Fig. 3** Sobel argument **Fig. 4** Sobel module

#### **4 Experiment Description**

Experiments have been performed with images from the DDSM and their preprocessing was the one described in the previous work by [5].

Lowest and highest bright values have been detected and readjusted using linear conversion to generate a grayscale image of 16 bits.

After this process, each image with shaped anomalies was applied the Sobel filter to generate both polar components.

Different back propagation neural network input setups had been tested including or excluding the mean, variance and size of each region in each of the three images (original, Sobel module and argument).

For each configuration, one third of the data had been kept back for the verification process. The remaining two thirds were used to train the network. After training the verification was performed running the network with the kept-back third of the data to compare the results with the previously known ones obtaining a success ratio.

## **5 Experimental Results**

After suppressing the error of each study considering it as a *True/False* answer and then calculating the success rate the achieved value was 73%, independently of its error interval. It was found a correlation between the wrong-detected images and bad-shaped tumors on the input database.

It was also observed that an argument Sobel filter and a variance were not necessary in any of the images and it has an overloading element for the network and, in some cases, adding noise and deteriorating the results.

The best results achieved had used 10 layers for each side of the boundary.

For the tested data set, the setup with a higher success rate was a neural network with three layers, having 12 neurons in the middle hidden one.

The average timing to process 271 images, including decompression and Sobel operator was of 42 minutes and the average training time was inside an interval between 3 a 5 minutes. The computer used was an average-fast computer sold in 2008.

## **6 Conclusions**

Stated workarounds achieved better success rates than the previous work, that is a 73% over a 60%.

The shape and boundary type of the tumors is important and the Sobel filter (only the module part) proved to be useful to detect them. The variance applied to all the images and the Sobel argument image had shown no improvement after the classification process.

Developing an algorithm to improve bound accuracy should increase dramatically the success rates.

## **References**

- 1. AMA, Consenso Nacional Inter-Sociedades sobre Cáncer de Mama: Pautas para el Diagnóstico y Manejo de las Lesiones Mamarias Subclínicas. Asociación Médica Argentina (2006)
- 2. Antonie, M., Zaïene, O., Coman, A.: Application of data mining techniques for medical image classification. In: Proceedings of the Second International Workshop on Multimedia Data Mining, San Francisco (2001)
- 3. Bellman, R.: On a Routing Problem. Quarterly of Applied Mathematics 16(1), 87–90 (1958)
- 4. Dijkstra, E.W.: A note on two problems in connexion with graphs. Numerische Mathematik 1, S269–S271 (1959)
- 5. Ferrero, G., Britos, P., García-Martínez, R.: Detection of Breast Lesions in Medical Digital Imaging Using Neural Networks. In: Debenham, J. (ed.) IFIP International Federation for Information Processing. Professional Practice in Artificial Intelligence, vol. 218, pp. 1–10. Springer, Boston (2006)
- 6. Heath, M., Bowyer, K., Kopans, D., Kegelmeyer, W.P., Moore, R., Chang, K., MunishKumaran, S.: Current status of the Digital Database for Screening Mammography. In: Digital Mammography. Proceedings of the Fourth International Workshop on Digital Mammography, pp. 457–460. Kluwer Academic Publishers, Dordrecht (1998)
- 7. Heath, M., Bowyer, K., Kopans, D., Moore, R., Kegelmeyer, W.P.: The Digital Database for Screening Mammography. In: Yaffe, M.J. (ed.) Proceedings of the Fifth International Workshop on Digital Mammography, pp. 212–218. Medical Physics Publishing (2001) ISBN 1-930524-00-5
- 8. Selman, S.: Data Mining of Digital Mammograms Will Aid in War against Cancer (2000), http://www.gatech.edu (accessed March 27, 2008)
- <span id="page-5-0"></span>9. Smith, R.A., Caleffi, M., Albert, U.S., Chen, T.H., Duffy, S.W., Franceschi, D., Nystrom, L.: Breast cancer in limited-resource countries: early detection and access to care. Breast J. 12(suppl. 1), S16–S26 (2006)
- 10. Sobel, I., Feldman, G.: A 3x3 Isotropic Gradient Operator for Image Processing. Presented at the Stanford Artificial Project (1968)
- 11. WHO. Screening for Breast Cancer (2008), http://www.who.int/cancer/detection/breastcancer/en/ index.html (accessed October 26, 2008)