Evaluation of hybrids algorithms for mass detection in digitalized mammograms

José Cordero1,2 and Johnson Garzón Reyes1

1Grupo de Óptica y Espectroscopía GOE. Centro de Ciencia Básica. Universidad Pontificia Bolivariana de Medellín. Colombia
2Grupo de Investigaciones en Bioingeniería GIBIOING. Centro de Bioingeniería. Universidad Pontificia Bolivariana Medellín. Colombia

josecorderog@hotmail.com

Abstract. The breast cancer remains being a significant public health problem, the early detection of the lesions can increase the success possibilities of the medical treatments. The mammography is an image modality effective to early diagnosis of abnormalities, where the medical image is obtained of the mammary gland with X-rays of low radiation, this allows detect a tumor or circumscribed mass between two to three years before that it was clinically palpable, and is the only method that until now achieved reducing the mortality by breast cancer. In this paper three hybrids algorithms for circumscribed mass detection on digitalized mammograms are evaluated. In the first stage correspond to a review of the enhancement and segmentation techniques used in the processing of the mammographic images. After a shape filtering was applied to the resulting regions. By mean of a Bayesian filter the survivors regions were processed, where the characteristics vector for the classifier was constructed with few measurements. Later, the implemented algorithms were evaluated by ROC curves, where 40 images were taken for the test, 20 normal images and 20 images with circumscribed lesions. Finally, the advantages and disadvantages in the correct detection of a lesion of every algorithm are discussed.

Key words: Mammographic lesions, ROC curves, Bayesian filter image, breast cancer.

1. Introduction

The breast cancer remains being a significant public health problem, and is the cause of many deaths among women [1]. An early detection of the lesions can increase the successful possibilities of some treatment [2]. The methods more used for the early diagnosis are: the breast self examination and the mammography [3]. The mammography is an effective imaging modality for diagnosing abnormalities [4], where the medical image is obtained of a mammary gland with X-rays of low radiation [5],[3],[6]. In this technique the tumors or circumscribed masses can be detected between two to three years before the clinical evidence. Additionally, is the only method where mortality by breast cancer is reduced about a 50% from a the human group defined [7]. For expert humans is difficult provide an accurate diagnosis [3], due to a variety of factors [1], like: poor quality of the image, benign
appearance of lesions, eye fatigue factor. However, the great difficulty is caused by the difference between the brightness zones of the objects on the mammograms; however, this difference (due to difference in density of tissues that absorbs the X-ray radiation) is the useful property of the mammographic techniques [2]. The more important signs that the human expert search in a mammogram, are the microcalcifications and masses, these can be the more important indicators of breast cancer [8]. Microcalcifications are clusters of calcium, of a size smaller than 1mm and often appear in the fatty tissue, these look in the mammogram like bright white spots [1]. On the other hand, a mass is more largest than a microcalcification, it can have different shape (circumscribed to spiculated), size (1mm to several cm), and often appear in the densest areas of the breast [3], therefore, are confused with the breast structure [1]. Hence the mass detection is more difficult that the microcalcifications detection. In this case, in very important the help of a Computer Aided Detection system (CAD system), for detecting the suspicious regions and focus the attention of the human expert onto a possible tumor [8],[1]. Basically a CAD system is a set of developed tools to attend the human expert in the detection and/or evaluation of mammographic images [9], these systems are used like pre-reading or second review methods [8], some studies estimated the sensibility of the radiologist without use a CAD system is 80%, while using a CAD system can get 90% of sensibility [3]. In this paper, different hybrids algorithms for circumscribed mass detection on digitalized mammograms are evaluated. In the first step correspond to a review of the enhancement and segmentation techniques and its application. After, a shape filtering was applied to the resulting regions. By mean of a Bayesian filter the survivors regions were processed, where the characteristics vector for the classifier was constructed with few measurements. Later, the implemented algorithms were evaluated by ROC curves, where 40 images were taken for the test, 20 normal images and 20 images with circumscribed lesions. Finally, the some conclusions are discussed

2. Characterization of the mammographic image

The most systems that digitized the mammographic films, have a spatial resolution of 50µm, with a range of gray levels of 12 bits [10], in this experimental work was taken the images of the mini-MIAS database [11], all the images have a spatial resolution of 200µm, have a gray scale level of 8 bits, the (x, y) coordinates of the lesions are provided by the database, this database have 322 images, previously classified, 207 images are normal cases and 115 remaining images have an abnormality. Figure 1 shows two mammographic images that have abnormalities, in some cases the abnormality is confused with the background of the breast, and the lesion is difficult establish, the tumors can recognized like low density locally areas in the mammograms, but not have absolute values and are not constants, the size vary, and the background and the brightness are different from one mammogram to another [8].

3. Proposed methods for detecting circumscribed masses

The processing of the mammographic image can be synthesized in five stages: normalization, enhancement, segmentation, feature extraction and classification. Every stage is explained in the next subsections. From each stage some algorithms are implemented forming three combinations for detecting circumscribed masses.
3.1 Normalization

The normalization is necessary by the mammograms may vary their gray level distribution and some can be more intense than others, due to breasts density. The normalization must be performed before the enhancement of the image, to allow the uniformity of intensity ranges in all mammograms, simplifying the diffusion, using only a single set of parameters [1]. In the studies realized by Cheng and Xu in [1], shows that the range of gray level of the lesions (masses and microcalcifications) is in 20 to 180, but applying a tolerance, the mammograms are normalized in 0 to 200, using the equation:

\[
x = \frac{\text{maxrango} \cdot (g_{\text{orig}} - g_{\text{min}})}{g_{\text{max}} - g_{\text{min}}}
\]

Where maxrango is the maximum intensity range of the normalized image, gmin and gmax are the gray levels mammogram minimum and maximum respectively. \(g_{\text{orig}}\) is the gray level before normalization, and \(x\) is the gray level after normalization process.

3.2 Enhancement techniques

The enhancement techniques are used for increase the differences between anomalies and the healthy tissue [3], with the aim of that human expert can observe better the details in the digital mammography [12], but the noise can increase while enhancement the contrast, or the fine details can be eliminated; this techniques are interactive methods, in the sense that the system of reference is the human vision [13].

3.2.1 Median filter

A median filter with a window of \(n \times n\) is applied for noise elimination, without converting the image to diffuse image and retaining its fine details [13]. The median filter operation can be implemented as the product of a constant \(k\) and a sequence \(f(j)\):
3.2.2 Gray scale transformation

The purpose of a gray level transformation is increasing the difference between bright and dark pixels, giving more relevance to the brightest pixels. By mean of a cubic transformation the image increase the difference between dark and bright regions [14]:

\[
\text{PL}(i, j) = \text{round}\left( P(i, j)^3 \frac{(i, j)^3}{(f \cdot M^2)} \right)
\]

\[ M = \max(P(i, j)) \]

Where \(P(i, j)\) is the original image, \(PL(i, j)\) is the resulting image, \(M\) is the maximum gray value of the pixels of the original image, and \(f\) is the brightener factor.

3.2.3 Linear transformation filter

This filter increases the contrast between several regions, and removes the signal noise. It can be implemented defining \(m\) as the maximum gray level, \(a\) and \(b\) as two positive real numbers, and giving a constant \(\alpha\) to the pixels of the original image \(OI(i, j)\) for producing a enhanced image \(EI(i, j)\), by the follow relationships:

\[
m = a \cdot \log(1 + m \cdot b)
\]

\[
b = \frac{1 - \exp(m / a)}{m}
\]

\[
\begin{align*}
\text{if } OI(i, j) < a, & \quad EI(i, j) = a \cdot \log[1 + b \cdot OI(i, j)] \\
\text{if} OI(i, j) > a, & \quad EI(i, j) = \frac{\exp(OI(i, j) / a) - 1}{b}
\end{align*}
\]

Empirically the values of \(\alpha = 10000\), and \(a = 0.3\), are chosen, Kom and Tiedeu in [8].

3.2.4 Bézier histogram

The Bézier histogram [15], is an smoothing histogram. It eliminates the zig-zag that have the original mammogram histogram, and helps to determine where is the center of the higher brightness level, to find a natural threshold and segmenting the original image. Qi and Snyder in [15], use 256 control points, equivalent to the brightness levels of the histogram, in the positions: \(p_k=(x_k, y_k)\), with \(k\) \(x_k\) varying from 0 to 255. Then the coordinate points \(p_k\) are blended to produce a position vector \(P(u)\), which describes the path of an approach Bézier between \(p_0\) and \(p_{255}\).
\[ P(u) = \sum_{k=0}^{255} p_k \cdot \text{BEZ}_{k, 255}(u) \]  

(9)

Where \( 0 \leq u \leq 1 \), \( \text{BEZ}_{k, 255}(u) \) are the Bernstein polynomials:

\[ \text{BEZ}_{k, 255}(u) = C(255, k) \cdot u^k \cdot (1-u)^{255-k} \]  

(10)

and \( C(255, k) \) are the binomial coefficients. Based on Bezier histogram the higher brightness levels can be localized, which is where the local minimum appears or inflexion point (located on the first and second derivative), see Fig. 2. This location is used as a threshold [16].

3.3 Segmentation techniques

The aim of the segmentation is partition an image in regions of interest that are homogeneous with respect to one or more features [12], in the mammographic context a segmentation algorithm is used for detecting the edges of the whole breast, or detect some abnormalities like masses or microcalcifications [9]. In next subsection some segmentation techniques are presented.

3.3.1 Binarization by adaptive local thresholding

This method was development by Kom and Tiedeu [8], for segmenting the areas of possible masses. The first step is to apply the linear transformation filter to the original image \( OI \) and the resulting image is subtracted of the original, obtaining an enhancement image in the bright areas,

\[ SI(i, j) = OI(i, j) - EI(i, j) \]  

(11)

Where \( SI(i, j) \) is the subtracted image and \( EI(i, j) \) is the enhancement image. The segmented image of the SI image, is carried out with an adaptive thresholding, using a rule for determine and classify if a
pixel belongs to a potential mass, or is a normal pixel:

\[
\text{if} \left( SI(i, j) \geq TH(i, j) \quad \land \quad SI_{\text{diff}} \geq MvoisiP \right)
\]

\[
TH(i, j) = MvoisiP + \gamma \cdot SI_{\text{diff}}
\]

\[
SI_{\text{diff}} = SI_{\text{max}} - SI_{\text{min}}
\]

Where \( SI(i, j) \) is an potential area (lesion), \( TH \) is the adaptive local thresholding, \( MvoisiP \) is an average of pixels intensity in a small window around \( SI(i, j) \), \( SI_{\text{max}}(i, j) \) and \( SI_{\text{min}}(i, j) \) are maximum and minimum intensity value in a large window, \( \gamma \) is a threshold coefficient that is choose empirically.

To obtain satisfactory results for producing a good segmentation of the region of interest (ROI), multiple values for the parameters \( a \) and \( \gamma \) are tested.

### 3.3.2 Region growing

The region growing methods seeking groups of pixels with uniform intensities [12], this approach begin with the “seed” points that are along the structures of interest, the seeds can be planted manually or automatically. The seed point \((s_x, s_y)\) is defined inside a possible mass like [17]:

\[
(s_x, s_y) \in L \quad \text{for all } L
\]

Where the perimeter of \( L \) should be closed. The next step is examining the connected pixels for the \( R \) region to grow, using a uniformity criterion [13],

\[
|r - S| < t \implies r \in R
\]

Where \( S \) is the seed point, \( r \) is the pixel to analyze, and \( t \) is a gray level. Then a region to grow if the difference of gray levels between \( S \) and \( r \) is less to \( t \) threshold.

### 3.3.3 Segmentation using k-means

Other type of segmentation technique can be performed using an algorithm of data mining “the \( k \)-means algorithm”, the finality of this grouping is recognize natural groups of data for producing a concise representation of a system behavior [18]. Before begin to grouping, the \( k \)-means algorithm requires set a number of \( k \) groups, is not hierarchy, the data partition is performed in \( k \) subsets. This grouping is basically determined by the objects distances [19]. This algorithm has the following structure: first, assign each object randomly to one of the clusters \( k=1, \ldots, K \). Second, compute the means of each of the clusters:

\[
\mu_k = \frac{1}{N_k} \sum_{z_i \in C_k} Z_i
\]

Third, reassign each object \( z_i \) to the cluster with the closest mean \( \mu_k \). Fourth, return to compute the
means, until the means of the clusters do not change anymore. Through testing, was established that \( k = 6 \) for the pooling of data, obtaining a good segmentation and computational performance.

3.4 Feature extraction

After the segmentation process, experimentally applied a shape filtering, this filter used shape and textural characteristics on each resulting object, to prevent deformed structures influencing the classifier. The measurements computed are: the area is of great importance, for discriminate outsize objects [20]. The area is defined as:

\[
A(S) = \sum_x \sum_y I(x,y) \Delta A
\]

(18)

Where \( I(x,y) = 1 \) if \((x,y)\) pixel belong to \(S\), zero other case. \(\Delta A\) is a pixel area, the area change with the scale changes.

The Perimeter defined as:

\[
P(S) = \sum_i \sqrt{(x_i - x_{i-1})^2 + (y_i - y_{i-1})^2}
\]

(19)

Where \(x(t) y(t)\) are a set of enclosed pixels in a \(S\) region [21].

The Circularity [20], is defined as:

\[
C(S) = \frac{4 \cdot \pi \cdot A}{P^2}
\]

(20)

Where \(A\) is the area, \(p\) is the perimeter of the object. The circularity measure the efficiency with which the limits of an object enclose an area [21], for a circular region the circularity is near 1.

The compactness [22], is calculated, as the division of the area by the perimeter:

\[
Co = \frac{P^2}{4 \pi A}
\]

(21)

The eccentricity [23] also is a circularity measure, while closer to zero is an circle, is defined as:

\[
Ec(S) = \frac{2 \cdot \sqrt{\left(\frac{\text{AxisMayor}}{2}\right)^2 - \left(\frac{\text{AxisMenor}}{2}\right)^2}}{\text{AxisMayor}}
\]

(22)

The Third moment, which is a texture measure, is a measure of the symmetry of the histogram, while closer to zero the distribution is symmetric [23], the third moment \( \mu_3 \) is defined as:
\[
\mu_j = m_j = \sum_{i=0}^{l-1} (z_i - m_i) \cdot p(z_i)
\] 

(23)

3.5 Classification techniques

The classification is essentially the heart of pattern recognition. The selection of features that will be used to distinguish between one kind and another is one of the most important steps of the classification process, once selected features, specimens should be obtained for features different classes. The classifier simpler to implement is the Bayesian classifier, this classifier is used to segment automatically the masses on a mammogram.

3.5.1 Bayesian classifier

From the objects that pass the filter, construct a characteristics vector d-dimensional \(x\), \(d = 1, 2, 3, \ldots\) to be classified by a Bayesian classifier [24], the classifier is previously trained with two class \(w_1 = \text{lesion}\) and \(w_2 = \text{normal tissue}\), \(w\) is consider like a variable described probabilistically. If \(P(w_1)\) is an a priori probability of a lesion from a detected region and \(P(w_2)\) is an a priori probability of a normal tissue and the sum of both classes is 1. Then, Bayesian decision rule can be implemented.

Rojas and Nandi [2], say that the probability density function generality is chosen a gaussian function, in this work was taken the gaussian function defined [23] as:

\[
p(x \mid w_j) = \frac{1}{(2\pi)^{d/2} |C_j|^{1/2}} e^{-1/2(x-m_j)^T C_j^{-1}(x-m_j)}
\] 

(24)

Where each density is completely specified by the coo-variance matrix \(C_j\) and the mean vector \(m_j\).

Whereby the Bayesian decision function for a class \(w_j\) is given by the form:

\[
d_j = \ln p(x \mid w_j) + \ln P(w_j)
\] 

(25)

Replacing the eq.(24) in the eq.(25):

\[
d_j = \ln P(w_j) - \frac{n}{2} \ln (2\pi) - \frac{1}{2} \ln |C_j| - \frac{1}{2} \left[(x-m_j^T C_j^{-1}(x-m_j)\right]
\] 

(26)

4. Implementation and evaluation results

Experimentally was combined the techniques revised previously and was formed three algorithms for detecting masses. In this work is assumed that the probabilities a priori for \(P(w_1)\) is 35.71% and for \(P(w_2)\) is 64.29%, this values was obtained when the database was analyzed.

Before construct the characteristics vector, the measures was analyzed qualitatively with scatterplots (Fig. 3), this graphics are useful to know that values are taking the characteristics of each class, and see how these values are grouped. Then each characteristic was analyzed and if the cluster is sufficiently discriminating like to establish two groups, the characteristic is valid and is used for
Figure 3. (a) Scatterplot area vs. circularity, the data values form clearly two groups. (b) Scatterplot uniformity vs. entropy, the data values are not discriminating as to form two groups.

4.1 Evaluation

The detection algorithms were evaluated basically in two aspects, the ROC (Receiver Operating Characteristic) curve [25] and the FROC (Free Response Operating Characteristic) curve [3], but too was calculated other parameter the “ideal point of the ROC curve” [13], the ideal point is calculated as in equ.(29). For each algorithm is calculated the specificity and sensibility, the data selected for the test was 40 images of the mini-MIAS database, 20 images was normal and 20 images have an abnormality, in the work of Carreras Cruz and Rayón Vilella in [14], are the 20 images with abnormalities used, so in this experimental work was taken the same images to have a point of comparison in the evaluation of the detected masses system.

\[
D = \sqrt{(x - 0)^2 + (y - 1)^2}
\]  

(27)

4.1.1 Results 1

In this algorithm were used the techniques of gray scale transformation and the linear transformation filter for enhancement of the image and for the segmentation was used the binarization by adaptive local thresholding. In this test, was obtained: a sensibility of 80.00%, a specificity of 90.00% and a ideal point of 0.24 and presented 3 false positive and 4 false negative cases. The Fig. 8a shows the ROC curve obtained to evaluate this algorithm.
Figure 4. (a) ROC curve, for a fraction of 0.1 of false positive cases there is an 80% of true positives. (b) Final result of the detection algorithm, the yellow mark indicates the lesion region.

4.1.2 Results 2

In this algorithm were used the techniques Bezier histogram and the region growing. In this test, was obtained: a sensibility of 50.00%, a specificity of 100.00% and an ideal point of 0.5 and presented 0 false positive and 10 false negative cases. The Fig. 9a shows the ROC curve obtained.

Figure 5. (a) ROC curve, for a fraction of 0 of false positive cases there is a 50% of true positives. (b) Final result of the detection algorithm, the yellow mark indicates the lesion region.

4.1.3 Results 3

In this algorithm were used the techniques median filter and convolution with a circular mask for enhancement the image and the k-means for the segmentation. In this test, was obtained: a sensibility of 85.00%, a specificity of 85.00% and an ideal point of 0.212 and presented 3 false positive and 3 false negative cases. The Fig. 10 shows the ROC curve obtained.
5. Conclusions

The employment of the enhancement and segmentation techniques heightens the perception of the anomalies in the mammographic images, and if not used an appropriated technique the other process are affected and the masses can not be detected correctly. The linear transformation filter and Bézier histogram perform better computing a region of interest and when the image background is of dense glandular tissue the details of the lesion are eliminated.

The region growing detect success the contour of the objects, however this algorithm have a disadvantage, and is that the seeds are put manually and their reproduction is difficult. In the binarization by adaptive local thresholding were made several test for set the parameter in $\gamma = 0.00009$. This technique rounding the object contour and missing the original contour. The shape filtering proposed let pass some small objects that have an area of 170 to 320 pixels, a compactness less of 1,17 and a circularity major to 0,85.

The best algorithm evaluated was the M+CCM+SK algorithm but have a big disadvantage that is the rounding contour of the object that is completely different of the original contour. In the evaluation of an algorithm while highest values are the sensibility and sensibility better is the classifier. The results of this paper could be improved employment other techniques of segmentation and using other type of characteristics like texture measures.

References

[1] Cheng H D and Huijuan Xu 2002 Information Sciences 148 pp 167-184
[2] Rojas D A and Nandi A K 2008 Computerized Medical Imaging and Graphics 32 pp 304-315
[3] Llobet A R 2006 Doctoral Thesis Aportaciones al Diagnóstico de Cáncer Asistido por Ordenador (Valencia: Universidad Politécnica) p 203
[4] Cascio D et al 2004 Nuclear Science Symposium Conference Record IEEE pp 2695-2699
[5] Strickland R N 2002 Image-Processing Techniques for Tumor Detection (New York: Marcel Dekker)
[6] Brandon M E and Villaseñor Y 2006 Cancerologia 1 pp 147-162
[7] Botero U J et al 2004 Obstetricia y Ginecología, texto integrado. 7ª edition (Bogota: QuebecorWorld)

[8] Kom G et al 2007 Computers in Biology and Medicine 37 pp 37 - 48

[9] Malagelada A O 2007 Automatic Mass Segmentation in Mammographic images Doctoral Thesis (Girona: Universitat de Girona) p190

[10] Astley S M and Gilbert F J 2004 Clinical Radiology 59 pp 390 - 399

[11] Suckling J et al 1994 International Congress Series 1069 pp 375-378

[12] Bankman Isaac N 2000 Handbook of Medical Imaging, processing and analysis 1ª Editon (San Diego: Academic Press)

[13] Forero Vargas M G 2002 Introducción al procesamiento digital de imágenes 1ª Edition (Bogotá: La Silueta Ediciones Ltda)

[14] Carreras Cruz M V and Vilella R. 2006 Electronics, Robotics and Automotive Mechanics Conference (Cuernavaca: IEEE Computer Society) Vol 1 pp 19-24

[15] Qi H. and Snyder W E 1998 Engineering in Medicine and Biology Society Vol 2 pp 1021-1024

[16] Leithold L. 1992 El Cálculo con Geometría Analítica. 6ª Edition (México: Harla)

[17] Kupinski M A and Giger M L 1998 IEEE Transactions on Medical Imaging 17 pp 510-517

[18] Suliga M et al 2008 Computerized Medical Imaging and Graphics 32 pp 502-512

[19] Van der Heijden F et al 2004 Classification, Parameter Estimation and State Estimation, An Engineering Approach using Matlab (England: John Willey & Sons)

[20] Patt W K 2001 Digital Image Processing PIKS Inside. 3ª Edition (California: John Wiley & Sons)

[21] Nixon M S and Aguado A S 2002 Feature Extraction and Image Processing 1ª Edition (Great Britain: Newnes)

[22] Braz Junior G et al 2006 Identificação de Massas em Mamografias usando Textura, Geometria e Algoritmos de Agrupamento e Classificação Workshop de Informática Médica - WIM VI pp 94-104

[23] Gonzalez E C and Woods R E. Eddins S L. 2004 Digital Image Processing Using Matlab. (New Jersey: Pearson Prentice Hall)

[24] Martinez W L and Martinez A R. 2002 Computational Statistics Handbook with Matlab (Florida: Chapman & Hall/CRC)

[25] Duda R O et al 2000 Pattern Classification 2ª Edition (Hardcover: Wiley-Interscience)