Comm. 3-6

Region and Boundary descriptor for analysis in breast cancer detection and diagnosis

Ali Cherif CHAABANI, Atef BOUJELBEN, Adel MAHFOUDI, Mohamed ABID.

CES-Computer, Electronic and Smart Engineering Systems Design Laboratory,
National School of Engineers of Sfax: Tunisia
http://www.ceslab.org/

alicherifchaabani@hotmail.fr, atef_boujelben@yahoo.fr, Adel.Mahfoudi@fss.rnu.tn ,
Mohamed.Abid@enis.rnu.tn

Abstract: The cancer treatment is currently effective only if it is detected at an early stage. In this state, this paper deals with the problem of analysis using feature extraction in digital mammogram. In fact, we evaluate the efficiency on boundary and region information possessed by mass region. We propose to compare result of a region and boundary features in analyzing Region Of Interested "ROI". The objective of this study is to ameliorate the diagnosis quality. We use the Digital Database for Screening Mammography "DDSM" for experiments. Some classifiers as Multilayer Perception "MLP" and K-Nearest Neighbors "kNN" are used to distinguish the pathological records from the healthy ones. Using KNN classifiers, we obtained 89, 74 % as sensitivity (percentage of pathological ROIs correctly classified) by RDM method. The results in terms of specificity (percentage of non-pathological ROIs correctly classified) grow around 94,50% using MLP classifier.

Keywords*: Boundary, Region, Features, Mammography, Analysis, Diagnosis.*

Résumé: Le traitement de cancer est efficace seulement s'il est détecté à un stade précoce. Ce papier traite le problème d'analyse afin d'avoir de l'aide à la décision et en particulier l'extraction des caractéristiques dans des régions d'intérêt. Dans ce travail, on propose de comparer les résultats obtenus par des vecteurs d'analyse de région et de contour. L'objectif de cette étude est d'améliorer la qualité du diagnostic. Pour tester notre travail, on a utilisé une base de données universelle DDSM (Digital Database for Screening Mammography) et deux classifieurs KNN (K-Nearest Neighbors) et MLP (Multilayer Perception). En utilisant le classifieur KNN, on obtient une sensitivité 89.74% par la méthode RDM « Radial Distance Mesure ». Les résultats en termes de spécificité peuvent atteindre 94.5% en utilisant le classifieur MLP.

Mots Clés : *Contour, Région, Caractéristique, Mammographie, Analyse, Diagnostic.*

1. Introduction

Breast cancer is considered as a major health problem that constitutes the strongest cause behind mortality among women in the world. However, although breast cancer incidence has increased over the last decade, breast cancer mortality has declined among women of all ages [1]. This favorable trend in mortality reduction may relate to improvements made in the breast cancer treatment and the widespread adoption of mammography screening. In the last decade many research efforts attempted a generalization of approaches used in general imaging processing to cope with a specific one; namely, medical image processing or analysis. In the past; several years ago, there was a tremendous evolution in mammography process. In processing techniques, every method based on segmentation and classification is adopted [2] [4]. However, breast tumors and masses appear in mammograms with different shape characteristics: malignant tumors usually have rough, microlobulated, or speculated contours; whereas benign masses commonly have smooth, round, macrolobulated, or oval contours. Measures that can quantitatively represent shape roughness and complexity can assist in the classification of malignant tumors and benign masses. The analysis technique is composed of two approaches: shape analysis [5] and texture analysis [6] [9] [10]. Shape analysis is composed of analysis of boundary and region. In the boundary analysis, the majority of researchers applied a method based on Radial Distance Measure "RDM", the convexity and the angular measure like Radial Angle [3] and Tuning Angle [9] [12]. In this context, Alvarenga et al [11] has evaluated the performance and relevance of seven shape features; namely perimeter "P", normalized radial value "NRV", standard deviation "SDEV", area ratio "AR", contour roughness "R", the circularity "C" and the morphological-closing ratio "Mshape". The performance of these features in distinguishing malignant from benign tumors reveals NRV as the best feature in terms of Az (0.91), sensitivity (88.2 %) and specificity (92.3%). AR has presented the highest sensitivity (93%); its specificity is the worst one (34.6%). These characteristics were enriched by adding some other various ones in [7]. The most important added characteristics are zero crossing "ZC" (i.e. a count of the number of times the radial distance plot crosses the average radial distance) and convexity "CVX", which allows representing the studied shape better than the characteristics cited above. Via these shape features, the authors have attempted a discrimination between malign and benign masses by using classification techniques. This work was resumed in [8] to carry out mammograms, segmentation and classification of identified ROI.

On the basis of this state of the art, we include the approach of shape analysis in our analysis process of mammograms. Shape feature extraction is composed of two approaches. The objective of this paper is to compare results of diagnosis adopted by feature based on boundary and region criteria.

The rest of this paper is organized as follows: section 2 describes the proposed scheme. Section 3 illustrates the method of analysis adopted. Section 4 presents the results obtained by the proposed scheme. Finally, we draw conclusions and some future issues in section 5.

2.The flow proposed in detecting breast cancer

The proposed scheme consists of three stages: identification of ROI, features extraction, and vector classification. Figure 1 shows the bloc diagram of the proposed scheme. ROI is selected from the image by fixing a rectangular box around the suspicious lesion area. A classical method of segmentation based on Sobel filter and seuillage is adopted. After the isolation of the ROI, features extraction is adopted on ROI: this is the stage of flow that raises a lot of interest in this paper. After that, a classification part is started. The features vector is entered to the classifier to make decision.

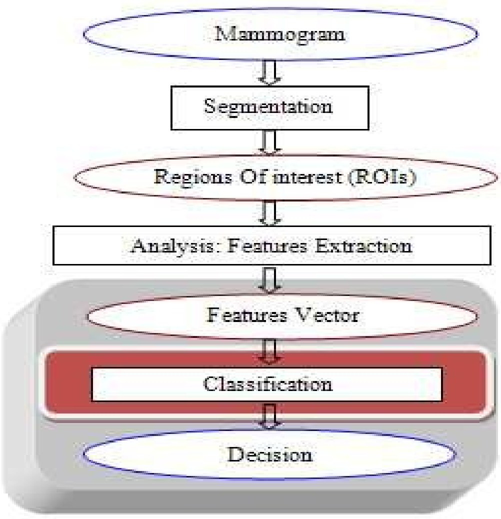

Figure 1: *Proposed Flow*

In the next section, we will illustrate the features extraction adopted in this paper.

3. Methods: Feature Extraction

Shape analysis is composed of analysis of boundary and region. In the following subsection we will illustrate the region analysis.

3.1 Region vector:

Region analysis which is often referred to as districting helps to define regions according to any criteria. We illustrate a method based on Circularity (C), Internal External Circle (IEC) and Normalized Residual Value (NRV).

A. Circularity: The circularity (C) describes the areas which can be circular. Circularity can be useful in this direction and can give an indication on the regularity of a given form. It is given by the following equation:

$$= \frac{4\prod}{} \qquad (1)$$

where P is the perimeter and A the area of the segmented tumour. The perimeter was measured by summing the number of pixels on the tumour contour, and the area was the number of pixels inside the contour.

B. Internal External Circle: The Internal External Circle (IEC) is near to 1 since the value of Inf_Radius is very close to the value of Sup_Radius, whereas for a lengthened form the value of IEC becomes close to 0 since the value of Inf_Radius is far from the value of Sup_Radius (figure 2).

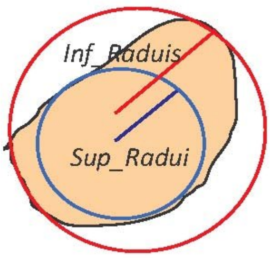

Figure 2: *Illustrative figure of internal and external circle.*

The IEC characteristic is given by the following equation: Inf _ Radius

$$= \frac{\_}{\_} \qquad (2)$$

where; *Inf_Radius*: the largest internal circle.
Sup_Radius: the smallest external circle.

The advantage of this characteristic is that it is invariant with any affine transformation. It is adequate with our work. Also, its calculation is slower, since for each form one must traverse all the points to determine the circle inscribed in the object which contains this point.

C. Normalized Residual Value: The Normalized Residual Value (NRV) is starting from the convex hull (Figure 3). NRV which gives best the performances compared to the other characteristics than have can extract, and which can be useful in the distinction between the regular and irregular areas.

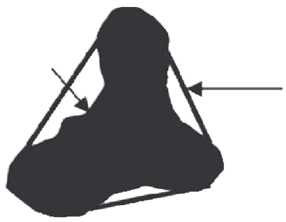

Figure 3: *Example of breast tumour and its respective Convex Hull.*

The NRV characteristic is given by the following equation: A region

$$= \frac{(\ \text{é}\ \ )}{(\ \ \ \_\ \ )} \qquad (3)$$

where: P is the perimeter of the convex hull and A is the area of the segmented tumour.

These three features represent the input of classifier. In the next section, we will show the performance of this shape vector in analysis of region in terms of diagnosis relevance.

3.2 Boundary vector

Boundary analysis which is often referred to as districting helps to define regions according to any criteria. In fact, Radial Distance Measure (RDM) is a famous method used in shape analysis.

We illustrate a method based on RDM in analyzing boundaries. However, an euclidian distance d(i) is calculated between the centroid C (gravity center) in region and all the points in boundary region (Figure4).

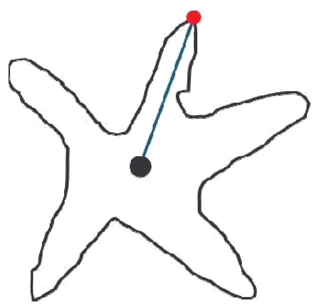

Figure 4: *Illustrative figure of RDM.*

$$(\ ) = \overline{(\ \ -\ \ ) + (\ \ -\ \ )}, 1 \leq \ \ \leq \qquad (4)$$

where (Xc,Yc) and (Xi,Yi)) are respectively the coordinates of the centroid C and the boundary pixel at the i-th location, N is the number of contour pixels. The coordinates of the centroid are given by:

$$= \frac{\sum}{\ } \qquad (5)$$

$$= \frac{\sum}{\ } \qquad (6)$$

where M: the number of points (pixels) in the area region.

To eliminate large calculations from characteristics, all radial distances d(i) are normalized by using the maximum value (normalised factor) of the radial distances:

$$(\ ) = \frac{(\ )}{[\ (\ )]} \quad (7)$$

From (7), three parameters were obtained: standard deviation (SDEV); area ratio (AR) and contour roughness (R). The features extracted in the RDM are cited below. We will only give the RDM features expressions related to this work.

A. Standard Deviation of the Normalized Radial Distance Measure (SDEV): This characteristic is defined as the variance of the distances $d_n(i)$ around the radius (the average radial distance measure) of a circle. SDEV gives better quality of information on the irregularity of contour. Indeed, when it is about a malignant tumour the value of SDEV tends towards a 0.5. On the other hand, in the case of benign tumour, the SDEV tends towards 0.

$$= \frac{\sqrt{\sum \quad (\ ) - \quad (\ )}}{} \quad (8)$$

where d (ı) is the mean value of d (i) and can be interpreted as the radius of a circular region.

B. Area ratio (AR): This characteristic computes the percentage of tumour outside the circular region defined by d (ı) (the average value of d (i)). The more irregular is the contour, the higher the value of AR. This characteristic discriminates between speculated and smoothed contours. It is illustrated in the following equation:

$$= \frac{1}{*\ (\ )} * \quad (\ ) - \quad (\ ) \quad (9)$$

where AR=0, if $(\ ) \leq \ (\ )$

C. Contour roughness (R): This characteristic treats micro-lobulated contours (contours which contain concave segments). It is defined as the average distance between neighbouring pixels over tumour contour. R is given by the following equation:

$$= \frac{1}{} \quad (\ ) - \quad (\ ) \quad (10)$$

Irregular contours provide high values of roughness index.

The advantage of these features is that it is standardized and it is invariant with any affine transformation. But, in practice, computation of these features increase the complexity of calculation. We are only interested in this paper in the quality of diagnosis, but not in the temporal complexity. In the following section, we will illustrate the results obtained in analysis region by two methods regions and boundary (RDM) features.

4. Results and discussion

The terminology used to determine the performance of a CADx (Computer Aided Diagnosis System) is defined as follows:

- *Sensitivity*: percentage of pathological ROIs correctly classified.
- *Specificity*: percentage of non-pathological ROIs correctly classified.

Because of the variation in the types of breast cancer, a larger number of cases can reduce the dependency "analysis techniques versus image sets". The performance of an algorithm is affected by the characteristics of a database like digitization techniques; namely pixel size, subtlety of cases, choice of training/testing subsets, etc.

4.1 The DDSM dataset:

The establishment of the DDSM allows the possibility of common training and testing data sets for the first time. The DDSM is the largest publicly available database of mammographic data. It contains approximately 2620 screening mammography cases. From the total number of ROI included in the DDSM database, we used 240 ROI malignant and 120 ROI benign. For the evaluation we used the remaining 260 ROIs that contain 130 ground troughed abnormal regions together with 130 entirely benign.

Their features represent the input of classifier. In the next section, we will show the performance of this shape vector in analysis of region in terms of diagnosis relevance.

4.2 Experimental results

The basic classification is based on the two methods of classification KNN and MLP as shown in Table 1 and Table 2: where Table 1 and Table2 show the results from discriminated analysis respectively in region and boundary analysis. Table 1 illustrates the importance of region information in shape analyzing. The result in terms of sensitivity tends towards 96% in MLP classifier. The result in terms of specificity tends towards more than 94% in case of this classifier. However, we can assume what the region vector is a good feature in differentiating the benign from the malign mass.

Table 1: *Results from region analysis*

classifier	KNN	MLP
sensitivity	95,45%	96,70%
specificity	92,96%	94,50%

Table 2 illustrates the importance of boundary information in shape analyzing in particularly in RDM approach. It also illustrates the importance of RDM in shape analyzing. The result in terms of sensitivity tends towards 89, 74 % in kNN classifier. The result in terms of specificity tends towards more than 85 % in case of the two classifiers kNN and MLP.

Table 2: *Results from discriminated analysis: RDM*

classifier	KNN	MLP
sensitivity	89,74%	86,88%
specificity	85,22%	85,43%

However, we can assume that the region vector is a good feature in differentiating the benign from the malign mass. In fact, the region vector is a better descriptor than the boundary one in terms of sensitivity and specificity. But, we should not conclude that it is the best or the worst because of the experimental condition. In fact, the digitization can reflect the final result.

5. Conclusion

In this work, we have attempted to improve the classification performance in shape approach in analyzing the mammographic images. We have presented a comparative study of region and boundary information. The results have been validated by two algorithms of classification: kNN and MLP. The results in terms of sensitivity and specificity are variable; they tend towards 94% and 96%. These results seem satisfactory, and the future work will be focused on ameliorating the performance of boundary features and also in the detection phase. The methods of detection like Level Set or wavelet will be considered. In the future work, we will also illustrate the effectiveness of combination of the statistical texture features and shape ones in diagnostic process.

6. References

[1] M. Lanyi.:Breast calcifications, Springer-Verlag, Berlin, Heidelberg, 1988.

[2]R. Amr Abdel-Dayem and R. Mahmoud El-Sakka.: Fuzzy Entropy Based Detection of Suspicious Masses in Digital Mammogram", In Images Proceedings of the 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference Shanghai, China, September 2005, pp 1-4.

[3] Sheng-Chih Yang, Chuin-Mu Wan, Yi-Nung Chung, Giu-Cheng Hsu, San-Kan Lee, Pau-Choo Chung and Chein-I Chang.: A Computer-Aided System for Mass Detection and Classification In Digitized Mammograms. In Biomed Eng Appl Basis Comm, 2005, pp 215-228.

[4] M. Sameti .: Detection of Soft Tissue Abnormalities in Mammographic Images for Early Diagnostis of Breast Cancer" University of British Columbia, November 1998.

[5]] U. Cascio, D. Fauci, B. Golosio, R. Magro, L.Masala, P. Oliva, G. Raso and S. Stumbo.: Massive Lesions Classification using Features based on Morphological Lesion Differences", In Transactions on engineering computing and technology, IMTC, March 2006.

[6] A. Boujelben, H. TMAR, J. M NIF and M. ABID.: Influence des caractéristiques issues de la matrice de co-occurrence sur le diagnostic du cancer du sein . In deuxième Workshop AMINA, Application Médicale en Informatique: Nouvelle Approche, Monastir, Tunisie, 17-19 Novembre 2006, pp 9-14.

[7] A. Retico, P. Delogu and M.E. Fantacci and P. Kasae.: An Automatic System to Discriminate Malignant from Benign Massive Lesions on Mammograms. In Medical Physics, volume 14, Jan 2007, pp 596-600.

[8] P. Delogu, M.E Fantacci, Kasae P.and A. Retico.: Characterization of mammographic masses using a gradient-based segmentation algorithm and a neural classifier.: In Comput Biol Med, 10 Mar 2008.

[9] Rangayyan, M. Rangaraj, D. Guliato, J. D. Carvalho and S. A Santiago.: Feature Extraction from the Turning Angle Function for the Classification of Contours of Breast Tumors. In The International Special Topics Conference on Information Technology in Biomedicine IEEE-ITAB 2006, pp 1-6.

[10] A.Oliver, X.Llado, R.Marti, J.Freixenet and R.Zwiggelaar. Classifying mammograms using texture information. In Medical Image Understanding and Analysis Aberystwyth and Wales and UK, July 2007, pp 223-227.

[11] A.V. Alvarenga, A. Fernando and C. I. Wagner Coelho de Albuquerque.: Assessing the performance of the normalised radial length and convex polygons in distinguishing breast tumors on ultrasound images.: Pereira Revista Brasileira de Engenharia Biomédica, volume 22, 2006, pp 181-189.

[12] G. Denise, M. Rangaraj, Rangayyan, D. Juliano Carvalho and A. Sérgio Santiago.: Polygonal Modeling of Contours of Breast Tumors With the Preservation of Spicules. In IEEE Transations In Biomedical Engineering, 2008, Volume 55.