Protuberance Selection descriptor for breast cancer diagnosis

Imen Cheikhrouhou, Khalifa Djemal, Hichem Maaref

▶ **To cite this version:**

Imen Cheikhrouhou, Khalifa Djemal, Hichem Maaref. Protuberance Selection descriptor for breast cancer diagnosis. 3rd European Workshop on Visual Information Processing (EUVIP 2011), Jul 2011, Paris, France. pp.280–285, 10.1109/EuVIP.2011.6045548 . hal-00654076

HAL Id: hal-00654076
https://hal.science/hal-00654076v1

Submitted on 18 Aug 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

PROTUBERANCE SELECTION DESCRIPTOR FOR BREAST CANCER DIAGNOSIS

Imene Cheikhrouhou, Khalifa Djemal and Hichem Maaref

IBISC Laboratory, Evry Val d'Essonne University
40 rue du Pelvoux, 91020, Evry, France
phone: + (33) 1 69 47 75 36, fax: + (33) 1 69 47 06 03
email: {imen.cheikhrouhou, Khalifa.Djemal, maaref}@iup.univ-evry.fr

ABSTRACT

In breast cancer field, researchers aim to automatically discriminate between benign and malignant masses in order to assist radiologists. In general, benign masses have smoothed contours, whereas, malignant tumors have spiculated boundaries. In this context, finding the adequate description remains a real challenge due to the complexity of mass boundaries. In this paper, we propose a novel shape descriptor named the Protuberance Selection (PS) based on depression and protuberance detection. This descriptor allows a good characterization of lobulations and spiculations in mass boundaries. Furthermore, it ensures invariance to geometric transformations. Experimental results show that the specified descriptor provides a promising classification performance. Also, results confirm that the new PS descriptor outperforms several shape features commonly used in breast cancer domain.

Index Terms— Computer aided analysis, Mammography, Shape analysis, Mass Protuberance Selection, Radial length features, Geometrical features, Curvature.

1. INTRODUCTION

Since several decades, breast cancer had regained a great importance from radiologists and researches. This kind of cancer threatens one of ten women life and the optimal way to decrease mortality rate is the early detection. In this context, mammography is widely recognized as the most reliable technique to perform such early detection. In mammographic images, benign masses appears with a circumscribed contour, whereas, malignant tumors are distinguished with spiculated boundaries.

Considerable works show the importance of shape features in breast cancer diagnosis [1]. In fact, the mass margin characteristics are the most important criteria deciding whether the mass is likely to be benign or malignant [2]. Several geometrical shape features where proposed in literature such as circularity, compactness and rectangularity [3]. Menut *et al.* [4] achieved a classification accuracy of 76% using four features based on mean and variance width of the parabolic segments. Also, Rangayyan *et al.* [5] propose an other boundary modeling method that treats the mass boundary as a union of piecewise continuous and locally-salient concave and convex parts. Authors achieved 91% as classification accuracy of circumscribed versus spiculated breast masses. Shi *et al.* [6] evaluate the classification performance of 24 individual features such as the spiculation and the patient informations. They demonstrate that, among these evaluated features, the spiculation measure had the best area under the Receiver Operating Characteristic (ROC) curve: $A_z = 0.78$. In fact, performed researches prove that features based on spiculation and concavity measures are very effective for mass characterization. Also, classification should be performed independently of the lesion position in the mammogram. Therefore, the shape description must be invariant in relation to geometric transformations such as translation, rotation and scaling.

In this context, we propose, the Protuberance Selection (PS) descriptor based on depression and protuberance detection for breast cancer classification. The PS descriptor, first, satisfies geometric invariance and, second, allows to well distinguish malignant from benign solid breast lesions. The remainder of the paper is organized as follows. Next section is preserved to detail the formulation of the proposed descriptor. Section 3 is allocated to experimental results. First, ROC curves are represented to assess the ability of the descriptor to discriminate between benign and malignant masses. Second, a comparison between PS and other descriptors is performed. The last section is dedicated to provide conclusions.

2. PROTUBERANCE SELECTION (PS)

Generally, a benign mass has a regular round or oval form with smoothed boundary. However, a malignant mass has an irregular form with a spiculated and a rough blurry boundary [7]. So, respecting a good shape characterization, improves at most mass classification performance. For this, we intend to detect lobulations that describe the spiculation rate of the mass contour. Therefore, we propose the Protuberance Selection (PS) as detailed below.

2.1. Spiculation detection

We follow the contour fluctuation by measuring the sign variation of the derivatives according to abscissa and ordinate. In fact, a derivative preserves the same sign when considering the contour in a given direction. Therefore, we could extract interest points allowing to characterize protuberances and depressions by detecting the derivative sign variation.

We consider the contour of a lesion C, defined on the interval I, as a set of p plane curves C_i in such manner that $C = \{C_1 \cup ...C_i \cup ...C_p\}$. Each plane curve C_i admits a parametric representation of class C^1 on the interval $I_i \in I$ so that for:

$$M(x,y) \in C_i \Leftrightarrow x = f(t), y = g(t) \tag{1}$$

where $t \in I_i$ and x and y are continuously differentiable on I_i. We denote by $M(t) = M(f(t), g(t))$ and we compute the derivatives $\frac{df^M(t)}{dt}$ and $\frac{dg^M(t)}{dt}$ of respectively $f(t)$ and $g(t)$ for each contour point M as follows:

$$\frac{df^M(t)}{dt} = \lim_{h \to 0} \frac{f(t+h) - f(t)}{h} \tag{2}$$

$$\frac{dg^M(t)}{dt} = \lim_{h \to 0} \frac{g(t+h) - g(t)}{h} \tag{3}$$

where $t + h \in I_i$ and $h > 0$. Since, the proposed measures of derivatives are sensitive to noise, we consider $h > 1$ which allows to smooth the contour and to obtain more stable derivatives.

We note by n the number of points in the contour and we compute initially, the n-dimensional vectors V_x and V_y. These vectors represent respectively the derivative of $f(t)$ and $g(t)$ with respect to t for each contour point M_k, $k \in \{1, 2, ..., n\}$. V_x and V_y equations are given by:

$$V_x = [\frac{df^{M_1}}{dt}, ..., \frac{df^{M_k}}{dt}, ..., \frac{df^{M_n}}{dt}] \tag{4}$$

$$V_y = [\frac{dg^{M_1}}{dt}, ..., \frac{dg^{M_k}}{dt}, ..., \frac{dg^{M_n}}{dt}] \tag{5}$$

The null values in V_x and V_y vectors represent stationary points obtained where the corresponding tangent is horizontal or vertical. We mention that when the second derivative is negative and the first derivative is null, we detect only inflection points and we could miss certain lobulations in the contour. So, we have just to follow the sign variation of the first derivative before and after stationary points. For this, we remove zero values from V_x and V_y. We define two new vectors V'_x of the size $n_1 \leq n$ and V'_y of the size $n_2 \leq n$ as:

$$V'_x = V_x \cap \Re^* \qquad \text{and} \qquad V'_y = V_y \cap \Re^* \tag{6}$$

where $\Re^*$ is the set of non null real numbers.

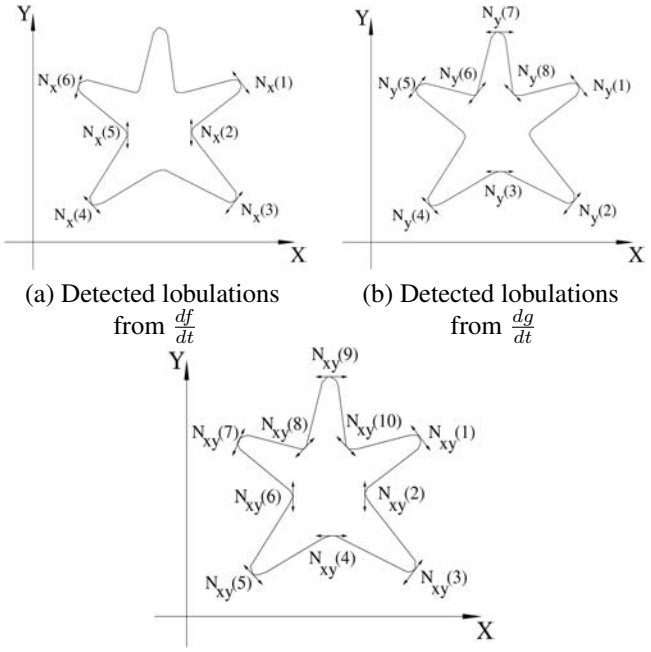
When two successive elements of V'_x (or V'_y) have the same sign, the contour keeps the same direction according to $f(t)$ (or $g(t)$). However, any sign variation between two successive elements implies a direction change and so a presence of lobulation. Let $signfollow$ be the indicator function allowing to follow the derivative sign variation:

$$signfollow(x) = \begin{cases} 1 \; if \; sign(x) \neq sign(x+1) \\ 0 \qquad\qquad otherwise \end{cases} \tag{7}$$

Using the indicator function $signfollow$, we determine the contour lobulation position. We store coordinates of detected lobulations from V'_x and V'_y in N_x and N_y matrices:

$$\begin{cases} N_x(i) = Coord(V'_x(k_x)), \; k_x \in \{1, 2, ..., n_1\} \\ s.t. \; signfollow(V'_x(k_x)) = 1 \end{cases} \tag{8}$$

$$\begin{cases} N_y(j) = Coord(V'_y(k_y)), \; k_y \in \{1, 2, ..., n_2\} \\ s.t. \; signfollow(V'_y(k_y)) = 1 \end{cases} \tag{9}$$



(a) Detected lobulations from $\frac{df}{dt}$

(b) Detected lobulations from $\frac{dg}{dt}$

(c) Lobulation detection of the lesion contour

Fig. 1. Protuberance and depression detection according to the derivative sign variation.

Considering that N_x and N_y sizes are respectively $(n_x, 2)$ and $(n_y, 2)$, we should note that $n_x \leq n_1$ and $n_y \leq n_2$. We define the matrix gathering the two sets of detected lobulation coordinates as $N_{xy} = N_x \cup N_y$ of the size $(n_{xy}, 2)$. Since the same lobulation could be detected twice through V'_x and V'_y sign variation, n_{xy} is always less or equal to $(n_x + n_y)$. Figure 1 illustrates interest points characterizing detected lobulations from $\frac{df}{dt}$ (figure 1.a) and $\frac{dg}{dt}$ (figure 1.b). Although the

three top points in figure 1.a have a derivative equal to zero, they are not considered since the derivative of $f(t)$ does not have any sign variation. It is also the case of points in the right and left side in figure 1.b. We neutralize missed point effect when superposing results obtained in the two figures. Figure 1.c demonstrates that N_{xy} covers all lobulations (protuberances and depressions) in the contour.

2.2. Protuberance selection

Computation of spiculation measure relies only on protuberances whereas detected lobulations include both protuberances and depressions. So, next operation is the elimination of depressions in order to keep only protuberances reflecting the number of spiculations. We exploit the fact that a protuberance is defined by at most 4 neighbors belonging to the lesion. We compute for each element in N_{xy} the intensity sum of its 8 neighbors. In fact, we consider $Neigh_i$ as the i^{th} neighbor of each element in N_{xy} and $Intensity(Neigh_i) = 1$ when the pixel is inside the lesion and $Intensity(Neigh_i) = 0$ when it is outside.

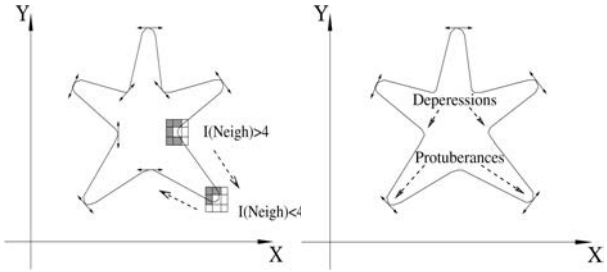


Fig. 2. a) Testing neighbors of lobulations, b) Protuberance selection.

We define Pr as a matrix containing coordinates of interest points characterizing the protuberances. In fact, Pr consists of N_{xy} elements which have at most 4 neighbors belonging to the lesion.

$$if \quad \sum_{i=1}^{8} Intensity(Neigh_i)(N_{xy}(i)) \leq 4 \quad \text{then} \quad Pr(j) = N_{xy}(i) \tag{10}$$

The Protuberance Selection PS based on depression and protuberance detection is then: $PS = length(Pr)$. Figure 2.a details the elimination of depressions by means of neighborhood intensity and figure 2.b summarizes the obtained protuberances.

3. CLASSIFICATION RESULTS

The general methodology of breast cancer Computer Aided Diagnosis (CAD) system contains principally three main steps namely segmentation, description and classification. The segmentation consists on extracting the mass contour from a region of interest. The lesion description uses specified features to characterize masses. Finally, classification allows to take decision and to distinguish between malignant and benign masses. In the following, we detail our adopted CAD system.

3.1. Used database

To perform benign versus malignant mass classification, first, images are selected from a publicly available database, the Digital Database for Screening Mammography (DDSM), assembled by a research group at the University of South Florida [8]. Mass boundaries in the DDSM have been subjectively characterized as (1) spiculated, (2) circumscribed, (3) ill defined, (4) microlobulated, and (5) obscured. The considered data set consists of 242 masses (128 benign/114 malignant) which are partitioned into 130 training (70 benign/60 malignant masses) and 112 test (58 benign/54 malignant masses) sets. Figure 3 shows some samples of the used data set. First line represents benign masses and second line malignant ones.

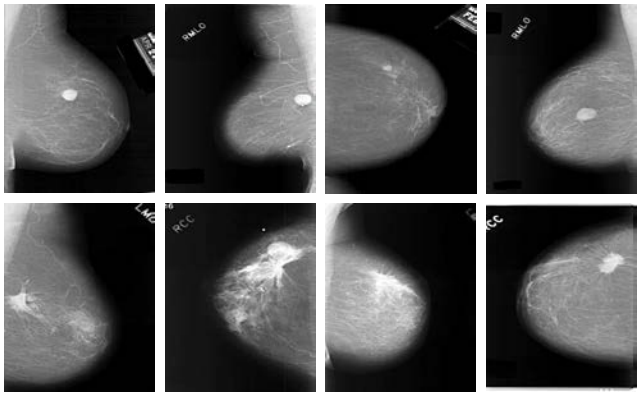
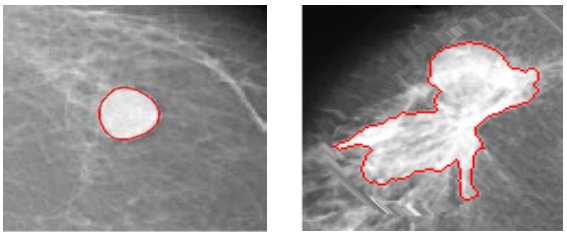


Fig. 3. Some samples of the used data set.

3.2. Segmentation

We define the boundary contour using the region-based active contour model as proposed by *Li et al.* [9]. The proposed model is able to segment images with intensity inhomogeneity. Also, it achieves good performance for images with weak object boundaries such the case of ill defined and obscured margins. We present in figure 4 two different masses segmented using the proposed level set evolution without reinitialization [9]. Figure 4.a shows a benign round mass with circumscribed margins while figure 4.b presents a malignant lobulated mass with microlobulated margins from DDSM database.



a) b)

Fig. 4. a) Case of a circumscribed circular mass and b) case of a lobulated mass.

3.3. Feature extraction

We compute the PS descriptor for the whole data set. In order to assess the pertinence of the proposed PS feature, we compare it to other shape features proposed in literature.

3.3.1. Normalized Radial Length features

Kilday *et al.* [10] developed a set of six shape features based on the Normalized Radial Length (NRL) from the objects centroid to the points on the boundary. The NRL features have had a good success in CAD applications and provide satisfying results [11]. Chen *et al.* [12] proposed new features from the NRL properties which had shown higher performance than basic NRL features. The normalized radial length $d(i)$ is filtered using a moving average filter and the filtered curve is noted $d_{ma}(i)$. From the NRL features proposed by Chen *et al.* [12], we select the difference of standard deviation (σ_{diff}) and the entropy of the difference between $d(i)$ and $d_{ma}(i)$ named E_{diff} defined as follows:

$$\sigma_{diff} = |\sigma - \sigma_{ma}| \tag{11}$$

where $\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(d(i) - d_{avg})^2}$ is the standard deviation of $d(i)$, σ_{ma} is the standard deviation of $d_{ma}(i)$ which is the result of $d(i)$ filtered using the moving average filter. As while tumor shape becomes more irregular, σ_{diff} approches higher values.

$$E_{diff} = \sum_{k=1}^{100} p_k log(p_k) \tag{12}$$

where p_k is the probability that $|d(i) - d_{ma}(i)|$ will be between $|d(i) - d_{ma}(i)|$ and $|d(i) - d_{ma}(i)| + 1/N_{bins}$. N_{bins} is the number of bins the normalized histogram, ranging in the [0,1] interval, has been divided in (N_{bins} = 100 in our analysis). This parameter is a measurement of the distribution for the difference between $d(i)$ and $d_{ma}(i)$. The E_{diff} value decreases while NRL approaches regularity.

3.3.2. Compactness

Among well known geometrical features, compactness noted Com has proven to be a good measure for classifying breast lesions by their shape [13].

$$Com = \frac{P^2}{A} \tag{13}$$

Where P and A are the mass perimeter and area respectively. Compactness represents the roughness of an objects boundary relative to its area. The smallest compactness values are for regular contours whose perimeters are lower than of complicated shapes.

3.3.3. Curvature

We compare, also, PS to the curvature which behavior is based on spiculations. In 2D, curvature at a given point A on curve is defined as the inverse of the radius of the osculating circle at A. The osculating circle can be found as follows: for any two points B and C near A compute the unique circle passing through A, B and C. If these points are collinear, than the circle has infinite radius and null curvature [14, 15].

$$Curv = \frac{1}{R} \tag{14}$$

The radius of the osculating circle is defined as:

$$R = \frac{a.b.c}{\sqrt{(a+b+c)(a-b+c)(a+b-c)(b-a+c)}} \tag{15}$$

where $a = |AB|$, $b = |BC|$ and $c = |AC|$.

In order to perform a reasonable comparison between the different features, we encode them on the same conditions (same database, segmentation method and classifier).

3.4. Classification and discussion

Finally, classification of breast masses is performed using the Support Vector Machine (SVM) classifier. SVM is a machine learning technique based on statistical learning theory [16]. When used in classification, SVM maps the input space to higher dimensional feature space and constructs a hyperplane, which separates class members from non-members. Given an input set X and a projection space F. The function that returns in the space F the inner product between two variables $x, x' \in X$ is known as the kernel function K. The SVM decision function is: $f(x) = \sum_{i=1}^{N} \alpha_i y_i K(x, x_i) + b$ where α and b are training parameters, x the query image.

We evaluate the classification performance by means of the Receiver Operating Characteristic (ROC) curve analysis using for this aim the same training and test database. Figure 5 shows the ROC curve resulting of PS classification by means of active contour segmentation and SVM classifier. The area under the ROC curve noted (A_z) which is the best

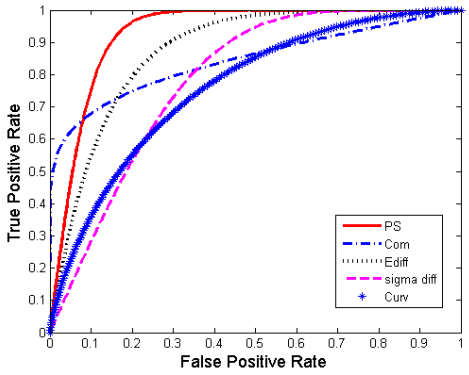


Fig. 5. Obtained ROC curves using PS, Com, E_{diff}, σ_{diff} and Curv descriptors.

Table 1. Individual performance of the different features in terms of the area under the ROC curve.

Feature	A_z
PS	0.93 ± 0.004
E_{diff}	0.87 ± 0.01
Com	0.84 ± 0.03
σ_{diff}	0.78 ± 0.05
$Curv$	0.76 ± 0.003

solution to evaluate a ROC curves is estimated to $A_z = 0.93$. The higher value of the area under the ROC curve proves that applied descriptor which preserves the same value of PS for the same form independently of its translation, rotation or scaling, insures a good classification rate.

Table 1 shows the performance of the different features in terms of the area under the ROC curve. We remark that the curvature relatively fails to well classify circumscribed/spiculated lesions with area under ROC $A_z = 0.76$. In such case, we have to determine, for each lesion, the corresponding threshold to find the correct number of spiculations. Such procedure prevents the feature extraction automation and also the use of an automated computer aided diagnosis system.

The difference of standard deviation (σ_{diff}) descriptor provides the area under ROC $A_z = 0.78$ while the entropy of the difference between $d(i)$ and $d_{ma}(i)$ noted E_{diff} provides satisfying results with $A_z = 0.87$. This result proves that the radial length measures could be used for the characterization of spiculations. In fact,NRL features proposed by [12] allow to characterize surface roughness. They are based on the centroid of a mass which provide satisfying results with a generally round boundary. However, in the case of complex shapes, the centroid may lie outside the tumor region and could not be a valid point to measure distance to the boundary.

The compactness Com provides a satisfying result with $A_z = 0.84$. It has been shown that geometrical features could be very informative and could improve classification results especially when they are used in junction with other features [17]. However, used individually they could not be robust enough to provide all necessary information about mass complexity.

The best result is obtained using the proposed Protuberance Selection descriptor. PS outperformes all the other shape-based features and seems to be the most effective in the benign versus malignant classification of breast masses. The proposed feature satisfies invariance criterion and is based on spiculation detection which is the most informative detail in contour about malignancy.

4. CONCLUSIONS

In this paper, we have proposed the Protuberance Selection (PS) descriptor based on depression and protuberance detection . This descriptor was devised to discriminate between benign and malignant masses in mammographic images in order to detect the breast cancer at its first stage. Classification efficiency is achieved using the area under the receiver operating characteristics (ROC) curve. Compared to known descriptors based on spiculation measures, PS provides better classification results and seems to be a relevant descriptor adapted to breast cancer recognition.

5. REFERENCES

[1] M. Mavroforakis, H. Georgiou, N. Dimitropoulos, D. Cavouras, and S. Theodoridis, "Significance analysis of qualitative mammographic features, using linear classifiers, neural networks and support vector machines.," *European Journal of Radiology*, vol. 54, pp. 80–89, 2005.

[2] C.J Vyborny, T. Doi, K.F. OShaughnessy, H.M. Romsdahl, A.C. Schneider, and A.A. Stein, "Breast cancer: importance of spiculation in computer-aided detection.," *Radiology*, vol. 215, pp. 703–707, 2000.

[3] B.S. Sahiner, H.P. Chan, N. Petrick, M.A. Helvie, and L.M. Hadjiiski, "Improvement of mammographic mass characterization using spiculation measures and morphological features.," *Med. Phys.*, vol. 28, pp. 1455–1465, 2001.

[4] O. Menut, R.M. Rangayyan, and J.E.L. Desautels, "Parabolic modeling and classification of breast tumours.," *Int. Journal of Shape Modeling*, vol. 3, pp. 155–166, 1997.

[5] R.M. Rangayyan, N.R. Mudigonda, and J.E.L. Desautels, "Boundary modelling and shape analysis methods for classification of mammographic masses.," *Medical & Biological Engineering & Computing.*, vol. 38, pp. 487–496, 2000.

[6] J. Shi, B. Sahiner, H.P. Chan, J. Ge, L. Hadjiiski, M.A. Helvie, A. Nees, Y.T. Wu, J. Wei, C. Zhou, Y. Zhang, and J. Cui, "Characterization of mammographic masses based on level set segmentation with new image features and patient information.," *Med Phys*, vol. 35, pp. 280–290, 2008.

[7] BIRADS, *American College of Radiology (BI-RADS) Breast Imaging Reporting and Data System*.

[8] M. Heath, K. Bowyer, D. Kopans, R. Moore, and P. Kegelmeyer Jr, "The digital database for screening mammography.," in *5th International Workshop on Digital Mammography, Toronto, Canada*, 2000.

[9] C. Li, C.Y. Kao, J.C. Gore, and Z. Ding, "Minimization of region-scalable fitting energy for image segmentation.," *IEEE Trans. Image Processing*, vol. 17, no. 10, pp. 1940–1949, 2008.

[10] J. Kilday, F. Palmieri, and M.D. Fox, "Classifying mammographic lesions using computer-aided image analysis.," *IEEE Trans. Med Imaging*, vol. 20, pp. 664–669, 1993.

[11] P. Delogu, M.E. Fantaccia, P. Kasae, and A. Retico, "Characterization of mammographic masses using a gradient-based segmentation algorithm and a neural classifier.," *Computers in Biology and Medicine*, vol. 37, pp. 1479–491, 2007.

[12] C.Y. Chen, H.J. Chiou, Y. Chiou Chou, H.K. Wang, S.Y. Chou, and H.K. Chiang, "Computer-aided diagnosis of soft tissue tumors on highresolution ultrasonography with geometrical and morphological features.," *Academic Radiology*, pp. 618–626, 2009.

[13] L. Shen, R.M. Rangayyan, and J.E.L. Desautels, "Application of shape analysis to mammographic calcifications.," *IEEE Trans. Med. Imag.*, vol. 13, no. 2, pp. 263–274, 1994.

[14] D. Coeurjolly, S. Miguet, and L. Tougne, "Discrete curvature based on osculating circle estimation.," *Proc. Int. workshop Visual Form.*, vol. 2059 of Lecture Notes in Computer Science, Springer, pp. 303–302, 2001.

[15] B. Kerautret, J.O. Lachaud, and B. Naegel, "Comparison of discrete curvature estimators and application to corner detection.," *Proceedings of ISVC'08: 4th International Symposium on Visual Computing*, vol. 5358 of Lecture Notes in Computer Science, Springer, pp. 710–719, 2008.

[16] N Vapnik, "An overview of statistical learning theory.," *IEEE Trans Neural Networks*, pp. 988–999, 1999.

[17] I. Cheikhrouhou, K. Djemal, D. Sellami, H. Maaref, and N. Derbel, "Empirical descriptors evaluation for mass malignity recognition.," *The First International Workshop on Medical Image Analysis and Description for Diagnosis Systems - MIAD'09*, pp. 91–100, 2009.