Mass Description for Breast Cancer Recognition

Imene Cheikhouhou, Khalifa Djemal, and Hichem Maaref

IBISC Laboratory, Evry Val d'Essonne University, 40 rue du Pelvoux, 91020, Evry, France
{Imen.CheikhRouhou,Khalifa.Djemal,maaref}@iup.univ-evry.fr

Abstract. In this paper, we present a robust shape descriptor named the Mass Descriptor based on Occurrence intersection coding (MDO) using the contour fluctuation detection. This descriptor allows a good characterization of the breast lesions and so a good classification performance. The efficiency of the proposed descriptor is evaluated on known Digital Database for Screening Mammography DDSM using the area under the Receiver Operating Characteristics (ROC) curve analysis. Results show that the specified descriptor has proven its performance in breast mass recognition using Support Vector Machine (SVM) classifier.

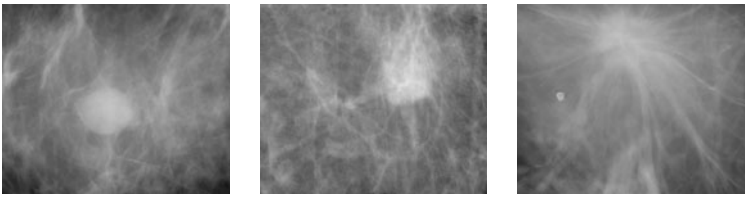
1 Introduction

Cancer is a disease characterized by an abnormal reproduction of cells, which invade and destroy adjacent tissues, being even able to spread to other parts of the body, through a process known as metastasis. Breast cancer is one of the major causes of deaths among woman. Presently, breast radiography, also called mammography, is the mostly used tool to detect this kind of cancer on its starting stage. The mammography makes possible the identification of the abnormalities in their initial development, which is a determining factor for success in treatment. According to BIRADS [1] standard, benign masses have round, smooth and well circumscribed boundary. In counter part, malignant masses have fuzzy, ill defined irregular and spiculated boundaries with spicules extending into surrounding tissues. However, certain benign entities such as fibroadenomas and cystic masses have also weakly defined boundaries. Similarly, malignant cases may have strong well defined boundaries (See Figure 1).

Several kinds of shape feature formulations were previously developed to characterize the mass boundaries. The most used descriptors are: circularity, rectangularity [2], compactness(C), spiculation index (SI), fractional concavity (F_{cc}) [3], fractal dimension [4], Fourier descriptors [5] and statistics based on the distribution of the Normalized Radial Length [6]. One important feature in automated malignity recognition is spiculation level characterization. This characterization should be able to regroup similar benign lesions in one class and similar malignant lesions in an other class without considering their position in the mammogram. Therefore, the shape description must be invariant in relation to geometric transformations such as translation, rotation and scaling.

There are many existing invariant texture analysis methods like Circular Mellin Feature extractor [11] and modified Gabor filters [9]. The circular mellin

A. Elmoataz et al. (Eds.): ICISP 2010, LNCS 6134, pp. 576–584, 2010.



a) b) c)

Fig. 1. Samples of mass margins extracted from BIRADS : a) circumscribed round, b) ill defined irregular and c) spiculated lesions

feature represents the spectral decomposition of the image scene in the polar log coordinate system and is invariant to both scale and orientation of the target texture pattern. In texture based pattern recognition, it is required to get rotation, scale and translation invariance, which would be insensitive to noise and illumination variation, but also in shape analysis to preserve the same value for similar shapes. In other terms, it is not judicious to obtain different classification outputs for objects having the same form but different positions, different orientations or different scales. Although the importance of shape descriptors invariance, few contributions were mentioned. Barman *et al* [8] proposed a shape measure which has a reasonable scale invariance. For this reason, we focus on this criteria to conceive an efficient and invariant shape descriptor able to quantify the degree of mass spiculation and to further improve the classification performance.

The paper is organized in four sections. Next one is preserved to detail the proposed descriptor and to provide examples clarifying the method. Section 3 shows experimental results. A ROC curve is represented to validate features ability to discriminate between benign masses and malignant tumors. We present also, a comparison with other methods that characterize shape complexity in the same data sets. Finally, section 4 is allocated to conclusion.

2 Mass Descriptor Based on Occurrence Intersection Coding (MDO)

Most of the shape analysis methods [5], [6], are focused on computing global measures characterizing the boundary's shape. Such methods are relatively insensitive to important local changes due to lobulations and spicules. While the majority of benign masses on mammograms are well circumscribed, some present stellate distortions. Also, while most malignant tumors are spiculated, some circumscribed malignant tumors are also encountered. In this way, discrimination between microlobulations in malignant tumors and macrolobulations in benign masses requires a detailed analysis of local characteristics of mass boundaries. Thus, we propose a new descriptor that could distinguish between benign and malignant lesions by detecting local changes in the mass boundary.

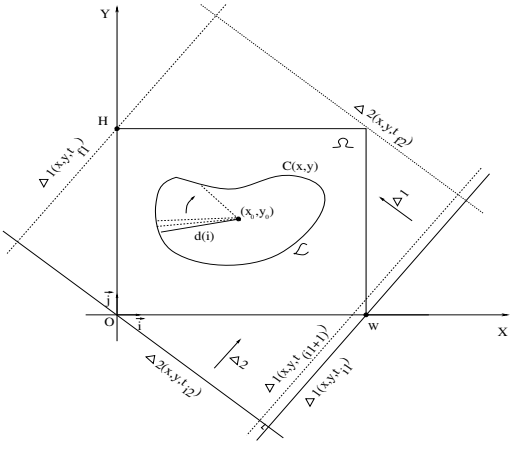


Fig. 2. Example illustrating the evolution of lines Δ_1 and Δ_2 in the domain Ω

2.1 MDO Principles and Computation

As illustrated in the figure 2, in the orthonormal basis $(o, \overrightarrow{i}, \overrightarrow{j})$, we consider an example of a lesion $\mathcal{L}$, in the domain Ω of the mammographic image which is delimited by the width w and the height h. Let $C(x, y)$ present the contour of $\mathcal{L}$ in Ω. Assume that n is the number of points in the contour of the lesion $\mathcal{L}$. We call $\Delta_1(x, y, t_1, \theta)$ the parallel lines chosen in a corresponding direction θ and has as equation:

$$\Delta_1(x, y, t_1, \theta) : \; b_1 y = a_1 x + c_1 + t_1. \tag{1}$$

Where $a_1, b_1, c_1, t_1, \theta \in \Re$. All different cases where a_1 and b_1 are not simultaneously null are taken into account in our procedure. When t_1 evolves in time, the line $\Delta_1(x, y, t_1, \theta)$ sweeps the domain Ω while preserving the same direction a_1. The major goal of this study is to calculate, in a given time t_1 and in a given angle θ, the number of intersections between the lesion and the line $\Delta_1(x, y, t_1, \theta)$. Similarly, we denote by $\Delta_2(x, y, t_2, \theta)$ the set of parallel lines chosen in a perpendicular direction $(\theta + 90°)$ to $\Delta_1(x, y, t_2, \theta)$ and has as equation:

$$\Delta_2(x, y, t_2, \theta) : \; b_2 y = a_2 x + c_2 + t_2 \tag{2}$$

Once the parameters of the lines Δ_1 and Δ_2 as the initial and final time required to sweep the whole domain Ω are set, we focus on computing intersections between the cited lines and the contour of the lesion. Thus, we calculate the vector S_1^θ of the size n_1 (respectively S_2^θ of the size n_2) where each element denotes the number of intersections between the contour $C(x, y)$ and the line $\Delta_1(x, y, t_1, \theta)$ at a specific time t_1 and a specific direction θ, (respectively the number of intersections between the contour $C(x, y)$ and the line $\Delta_2(x, y, t_2, \theta)$ at a specific time t_2 and a specific direction $(\theta + 90°)$.

$$\begin{cases} S_1^\theta(t_1) = \Delta_1(x, y, t_1, \theta) \cap C(x, y), & t_{i1} \leq t_1 \leq t_{f1} \\ S_2^\theta(t_2) = \Delta_2(x, y, t_2, \theta) \cap C(x, y), & t_{i2} \leq t_2 \leq t_{f2} \end{cases} \tag{3}$$

The elements of the vectors S_1^θ and S_2^θ depends on the complexity of the contour. In fact, the number of intersections between the contour function and each line is around the values $(2, 3, 4)$ if the contour function delineates a circumscribed circular or oval lesion. Homologously, the intersection between both functions could be of highly values (6,8,10,...) if the contour is more irregular and presents more lobulations or spiculations (as represented in Figures 3 and 4). In such manner, if we sum the elements in both vectors S_1^θ and S_2^θ, we would obtain low entities for regular lesions and high entities for irregular forms.

Nevertheless, the mass size factor could affect the significance of the measure, since a regular large size mass could reach a higher value than that of a small size spiculated mass. So, when summing the elements of both vectors S_1^θ and S_2^θ, obtained entity is non invariant to scaling and depends hardly on the size of the lesion. For instance, the same lesion considered in two different scales necessarily provides different results. In order to satisfy scaling invariance, we proceed by preserving only the variations of the topology which does not depend on the lesion size. We compute the vectors s_1^θ and s_2^θ (respectively of the sizes m and n) deriving from S_1^θ and S_2^θ, where each element points to the localization of the topology changing as follows:

$$s_1^\theta(m) = S_1^\theta(t_1) \quad \text{if} \quad g_1(t_1).S_1^\theta(t_1) \neq 0, \quad m \leq n_1 \tag{4}$$

$$s_2^\theta(n) = S_2^\theta(t_2) \quad \text{if} \quad g_2(t_2).S_2^\theta(t_2) \neq 0, \quad n \leq n_2 \tag{5}$$

$$\text{where } g_k(t) = \begin{cases} 1 & \text{if} \quad S_k^\theta(t) \neq S_k^\theta(t+1), \qquad k = 1, 2 \\ 0 & \quad otherwise \end{cases} \tag{6}$$

The mass descriptor computed for a precise angle θ results of the sum of elements of s_1^θ and s_2^θ which is independent of the considered scale.

$$MDO = \sum_{i=1}^{m} s_1^\theta(i) + \sum_{j=1}^{n} s_2^\theta(j) \tag{7}$$

2.2 Evaluation of the MDO Descriptor

To more explain the used terms, we study the case of two different lesions. Lesion 1 as represented in Figure.3 has a slightly spiculated form, whereas, lesion 2 (Figure 4) has a highly spiculated form. In order to simplify the representation and to more explicit results to reader, we consider the angle θ=0° which provides horizontal lines for the set $\Delta_1(x, y, t_1, 0°)$ and vertical lines for the set $\Delta_2(x, y, t_2, 90°)$. First column of Figure 3 shows the occurrence number $S_1^{0°}/S_2^{0°}$ of points belonging to Δ_1/Δ_2 and the contour simultaneously. Second column depicts the topology changing $s_1^{0°}$ and $s_2^{0°}$ of $S_1^{0°}$ and $S_2^{0°}$. For any size

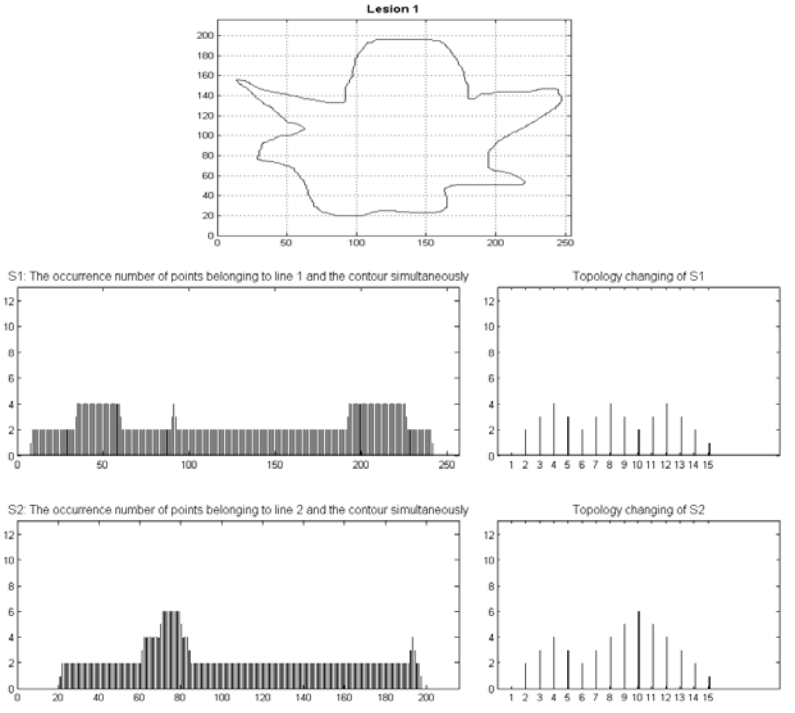


Fig. 3. MDO computation for the lesion 1: First column: The occurrence number (S1) of points belonging to line 1 and the contour simultaneously and the occurrence number (S2) of points belonging to line 2 and the contour simultaneously. Second column: s_1^{θ} and s_2^{θ} the topology changing of S1 and S2.

of this object, $S_1^{0^\circ}$ and $S_2^{0^\circ}$ would have the same template and would be smaller or greater depending on scaling level. However, $s_1^{0^\circ}$ and $s_2^{0^\circ}$ would preserve the same sequence and the same elements. We remark that, in the case of lesion 1, $S_1^{0^\circ}$ elements do not overpass the value 4 and $S_2^{0^\circ}$ elements do not overpass the value 6. Also, $s_1^{0^\circ} = 39$ and $s_2^{0^\circ} = 47$ which results of $MDO^{0^\circ} = 86$. These values are relatively low when compared to values obtained with the lesion 2. As represented in Figure 4, $S_1^{0^\circ}$ reaches the value 9 and $S_2^{0^\circ}$ reaches the highest value of intersection points in a given position of Δ_2 equal to 12. These important values lead to higher values of $s_1^{0^\circ} = 239$ and $s_2^{0^\circ} = 361$ which provides a higher $MDO^{0^\circ} = 600$. These results show that the proposed descriptor MDO is robust and could well distinguish between smoothed and spiculated contours.

3 Experimental Results

The proposed descriptor is evaluated through a publicly available database, the Digital Database for Screening Mammography (DDSM), assembled by a research group at the University of South Florida [12]. In this study, we select only regions

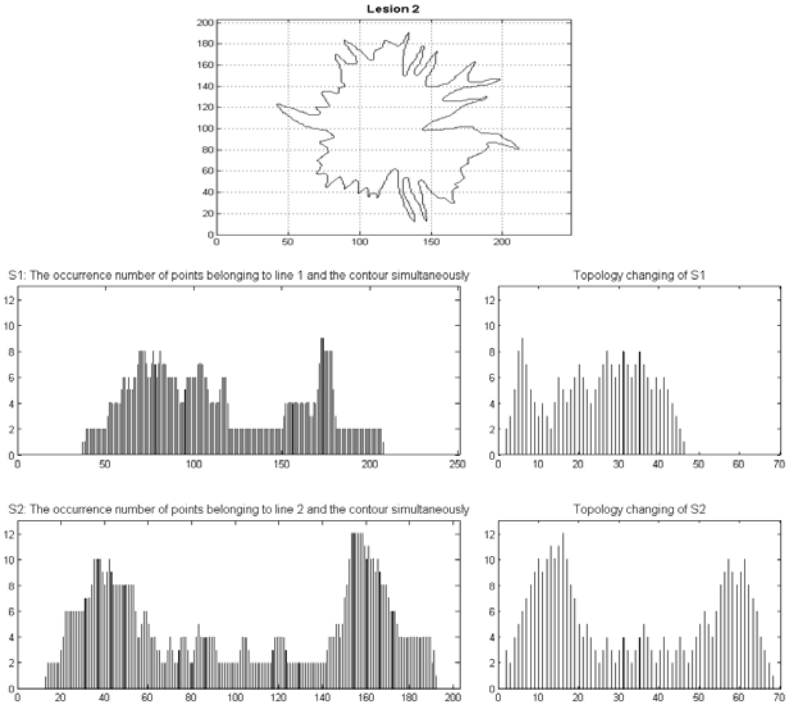


Fig. 4. MDO computation for the lesion 2: First column: The occurrence number (S1) of points belonging to line 1 and the contour simultaneously and the occurrence number (S2) of points belonging to line 2 and the contour simultaneously. Second column: s_1^θ and s_2^θ the topology changing of S1 and S2

of interests that depicted circumscribed masses or microlobulated/spiculated masses. The considered data set consists of 87 masses (49 benign/38 malignant) which are partitioned into training (29 benign/21 malignant) and test (20 benign/17 malignant) sets. We used as classifier the support vector machine SVM which is a well known machine learning technique based on statistical learning theory [10]. To evaluate the classification performance, we use the so-called Receiver Operating Characteristic (ROC) curve analysis, which is now routinely used for many classification tasks. It is represented by the classification sensitivity (TPF) as the ordinate versus the specificity (FPF) as the abscissa. These are defined as: $TPF = \frac{TP}{TP+FN}$ and $FPF = \frac{FP}{FP+TN}$ [7]. A good classifier has its ROC curve climbing rapidly towards upper left hand corner of the graph. This can also be quantified by measuring the area under the (ROC) curve: (A_z). Therefore, the closer the area is to 1, the better the classifier is, while the closer the area is to 0.5, the worse the classifier is.

We compare results from MDO to both Radial length and geometrical features. Kilday [6] developed a set of shape features based on radial length from the objects centroid to the points on the boundary. These features have had a

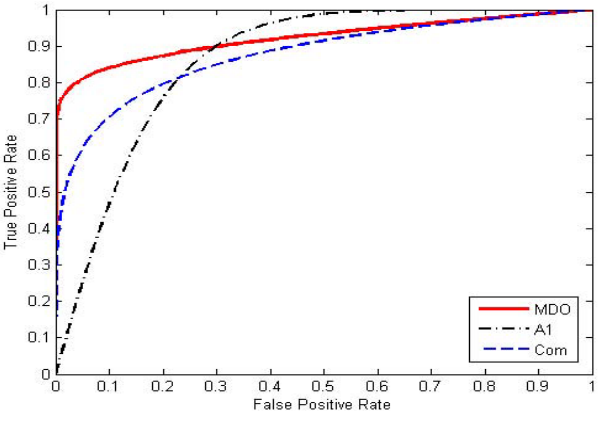


Fig. 5. ROC curves obtained with SVM classifier: Comparison between MDO, NRL (A1) and geometrical (Com) features

good success in computer aided diagnosis applications as demonstrated in [13] [14]. The mean of the normalized radial length is defined as the Euclidean distance from the center of mass of the segmented lesion to the i^{th} pixel on the mass boundary and normalized with respect to the maximum distance found for that mass. Figure 2 shows the mass center and how to calculate the radial distance. The area ratio A was a measure of the percentage of tumor outside the circular region defined by the mean of the x-y line plot, and alone described the macroscopic shape of the tumor. Since the area ratio have proven its performance in many applications [13], we follow the behavior of this descriptor applied to the same database in Figure 5. We also assess the performance of a geometrical feature: the compactness in differentiating between benign and malignant masses. The compactness (Com) is a measure of contour complexity versus enclosed area, defined as: $Com = \frac{P^2}{A}$. Where P and A are the mass perimeter and area respectively. A mass with a rough contour will have a higher compactness than a mass with a smooth boundary [15].

Considering results on Table 1 and Figure 5, first, we remark that all of the features, could effectively discriminate circumscribed benign masses from spiculated malignant tumors. Second, considering each feature individually, we notice that the area A relatively failed to well classify circumscribed/spiculated lesions and provide an area under ROC A_z=0.86. In counter part, the compactness Com provides satisfying result with A_z=0.90. The feature MDO of the present study has clearly outperformed all the other shape-based features and seems to be the most effective in benign/malignant classification of beast masses.

As demonstrated in last works, geometrical features could be very informative and could improve classification results especially when they are used in junction with other features [16]. However, used individually they could not provide all necessary information about mass malignancy. Regarding Radial length

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

	Tested features	*Corresponding* A_z
Proposed descriptor	MDO	0.92 ± 0.01
Geometrical feature	Com	0.90 ± 0.01
NRL feature	A	0.86 ± 0.03

measures, although they have proven their efficiency in many applications [13], we note that these features 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 the distances to the boundary.

The highest value of area under ROC which is obtained with MDO measure and presented in Figure 5 as a red line, is estimated to A_z=0.92. The higher value of A_z proves the efficiency and the robustness of the proposed descriptor. Also, this higher value of the area under the ROC curve proves that applied descriptor which preserves the same value of MDO for the same form independently of its translation, rotation and scaling, insures a good classification rate. This result proves that the proposed Mass Descriptor based on Occurrence intersection coding (MDO) could be a determining descriptor in computer aided diagnosis systems providing a second opinion to radiologists.

4 Conclusion

In this paper, we proposed a new descriptor dedicated to highlight spiculation measure in the analysis of breast masses. For evaluation, we have used the well known DDSM database and the SVM classifier. When computing the Mass Descriptor based on Occurrence intersection coding (MDO), we notice its ability to capture diagnostically important details of shape related to spicules and lobulations. The proposed descriptor provides high classification accuracies while discriminating between benign breast masses and malignant tumors. In the future work, we intend to study the MDO invariance in different rotation cases of the pathological masses.

References

1. American College of Radiology BI-RADS (Breast Imaging Reporting and Data System) Frensh Edition realized by SFR, 3rd edn. (2003)
2. Sahiner, B.S., Chan, H.P., Petrick, N., Helvie, M.A., Hadjiiski, L.M.: Improvement of mammographic mass characterization using spiculation measures and morphological features. Med. Phys. 28(7), 1455–1465 (2001)
3. Rangayyan, R.M., Mudigonda, N.R., Desautels, J.E.L.: Boundary modelling and shape analysis methods for classification of mammographic masses. Med. Biol. Eng. Comput. 38, 487–496 (2000)

4. Rangayyan, R.M., Nguyen, T.M.: Fractal Analysis of Contours of Breast Masses in Mammograms. Journal of Digital Imaging 20(3), 223–237 (2007)
5. El-Faramawy, N.M., Rangayan, R.M., Desautels, J.E.L., Alim, O.A.: Shape factors for analysis of breast tumors in mammograms. In: Canadian Conference on Electrical and Computer Engineering, pp. 355–358 (1996)
6. Kilday, J., Palmieri, F., Fox, M.D.: Classifying mammographic lesions using computer-aided image analysis. IEEE Trans. Med. Imaging 20, 664–669 (1993)
7. Adler, W., Lausen, B.: Bootstrap estimated true and false positive rates and ROC curve. Journal of Computational Statistics and Data Analysis, 1–12 (2008)
8. Barman, H., Granlund, G., Haglund, L.: Feature extraction for computer aided analysis of mammograms. In: International Journal of Pattern Recognition and Artificial Intelligence (IJPRAI), pp. 1339–1356 (1993)
9. Sim, D.G., Kim, H.K., oh, D.I.: Translation Scale and Rotation Invariant Texture Descriptor for Texture based Image Retrieval. In: International Conference on Image processing proceedings, vol. 3, pp. 742–745 (2000)
10. Vapnik, V.: Statistical learnig theory. Wiley, New York (1998)
11. Vyas, V.S., Rege, P.P.: Radon Transform application for rotation invariant texture analysis using Gabor filters. In: Proceedings of NCC-2007, IIT Kanpur, pp. 439–442 (2007)
12. Heath, M., Bowyer, K., Kopans, D., Moore, R., Kegelmeyer, P.: The Digital Database for Screening Mammography. In: 5th International Workshop on Digital Mammography, Toronto, Canada (2000)
13. Chen, C.Y., Chiou, H.J., Chou, Y.C., Wang, H.K., Chou, S.Y., Chiang, H.K.: Computer-aided Diagnosis of Soft Tissue Tumors on High-resolution Ultrasonography with Geometrical and Morphological Features. Academic Radiology, 618–626 (2009)
14. Hadjiiski, L., Chan, H.P., Sahiner, B., Helvie, M.A., Roubidoux, M.A., Blane, C., Paramagul, C., Petrick, M.N., Bailey, J., Klein, K., Foster, M., Patterson, S., Adler, A., Nees, A., Shen, J.: Improvement in Radiologists Characterization of Malignant and Benign Breast Masses on Serial Mammograms with Computer-aided Diagnosis: An ROC Study. Radiology, 255–265 (2004)
15. Shen, L., Rangayyan, R.M., Desautels, J.E.L.: Detection and classification of mammographic calcifications. Intemational Journal of Pattern Recognition and Artificial Intelligence 7(6), 1403–1416 (1993)
16. Cheikhrouhou, I., Djemal, K., Sellami, D., Maaref, H., Derbel, N.: Empirical Descriptors Evaluation for Mass Malignity Recognition. In: The First International Workshop on Medical Image Analysis and Description for Diagnosis Systems - MIAD'09, Porto, Portugal, January 16-17, pp. 91–100 (2009)