

Mass Segmentation using a Pattern Matching Approach with a Mutual Information Based Metric

Arnau Oliver ¹, Jordi Freixenet, Joan Martí, Marta Peracaula,
Computer Vision and Robotics Group, University of Girona, Girona, Spain

Abstract. As an ongoing effort to develop a computer aided system for the detection of masses on mammograms, we propose on this work a new model-based segmentation algorithm. The algorithm is based on a template matching scheme by using the mutual information approach in the similarity metric. Thus, the system will be able to determine if it exists a true mass on the studying image. The proposal was developed and evaluated using a database of 120 mammograms, 40 mammograms with confirmed masses and 80 normal ones. CAD performance was assessed using Receiver Operating Characteristics (ROC) and Free Receiver Operating Characteristics (FROC) analysis. The results prove the validity of the proposed method.

Keywords. Mammography, Image Segmentation, Template Matching

1. Introduction

Breast cancer is considered a major health problem in western countries, and constitutes the most common cancer among women in the European Union [4]. A study developed in 1998 by the American Cancer Society estimates that in western cultures between one in eight and one in twelve women will develop breast cancer during their lifetime [1]. Mammography remains the key screening tool for breast abnormalities detection, because it allows identification of tumour before being palpable. However, of all lesions previously diagnosed as suspicious and sent for biopsy, approximately 25% were confirmed malignant lesions, and approximately 75% were diagnosed benign. This high false-positive rate is related with the difficulty of obtaining accurate diagnosis [2]. In this sense, computerized image analysis is going to play an important role in improving the issued diagnosis. Thus, Computer-Aided Diagnosis (CAD) systems are composed of a set of automatic and semiautomatic tools to help radiologists to detect and diagnose new cases.

Although there are a large number of different types of mammographic abnormalities [7], this work is focused on mass detection. Breast mass is a generic term to indicate a localized swelling, protuberance, or lump in the breast, which can be caused by differ-

¹Correspondence to: Arnau Oliver, Computer Vision Lab, Universitat de Girona, Campus Montilivi, Ed P-IV. Av Lluís Santaló, sn. 17071, Girona. Catalunya, Spain. Tel.: +34 972 418891; Fax: +34 972 418259; E-mail: aoliver@eia.udg.es.

ent processes: from natural changes to cancerous processes. Masses are characterized by their location, size, shape, margin, and associated findings (i.e. architectural distortions, contrast, etc). As stated Li et al. [10], masses are one of the most difficult mammographic lesions to detect, due to:

- Masses are of varying size, shape and density.
- Masses often exhibit poor image contrast.
- Masses may be highly connected to surrounding parenchymal tissue, specially spiculated ones.
- Background tissues surrounding the mass are non-uniform and have characteristics similar to the mass.

Many methods have been developed for the automatic detection of masses in mammograms. Representative works were the proposal of Brzakovic et al. [3], which was based on a multi-resolution fuzzy pyramid linking approach, and the one of Sahiner et al. [14], who used k-Means clustering algorithm for grouping those pixels with similar grey-level values. Instead of use directly the pixels of the image to find the lesions, other proposals are based on an initial preprocessing of the image for enhancing suspicious regions, in order to make easy their posterior detection. For instance, the work of Petrick et al. [12] was based on a filtering approach in order to enhance the mass. In contrast, the approach of Kobatake et al. [6] was based on the detection of linear structures representing spicules. Zwiggelaar et al. [18] used both strategies in a two-directed algorithm: one direction looks for the center of masses, while the other looks for possible spicules. The union of both directions was hypothesized to be a spiculated mass. A different approach was proposed by Karssemeijer and te Brake [5], who used a pattern recognition scheme: first a set of features were extracted from the mammogram and, later, these were classified using two different classifiers. Note that a set of mammograms with ground-truth have to be used for their training.

However, note that those above approaches do not use shape information in the segmentation process. This is a general trend in mammographic mass segmentation, where most of the approaches are based only on grey-level and gradient features, while only few works use shape features. The main reason for such miss-using is due to the fact that masses can appear in a great range of shapes and sizes. The works that used shape features can be gather into two groups: based on region growing techniques or based on a pattern matching approach. Example works of the first approach are the ones of Kupinski and Giger [8] and Zheng et al. [17]. The former developed two extended region growing techniques, one based on gradient information and the other using a probabilistic model. In both approaches, the shape was used in the homogeneity criterion. In a similar way, the latter used gradient and shape information to constraint the region growth.

On the other hand, the segmentation of mammographic images by using a pattern matching approach was first proposed by Lai et al. [9], in one of the eldest works of mass segmentation. In their work, they look for circular masses by using a circular template and calculating the similarity between the template and each pixel of the mammogram using the normalized cross-correlation measure. The drawback of such approach was that the large variation on the sizes of the masses makes the algorithm slow. Moreover, it cannot be used to detect masses with spiculated shapes. This way, such scheme were in disuse until a recent work of Tourassi et al. [16], who reused this strategy to determine if a query Region of Interest (RoI) depicted a true mass. A RoI is an image containing

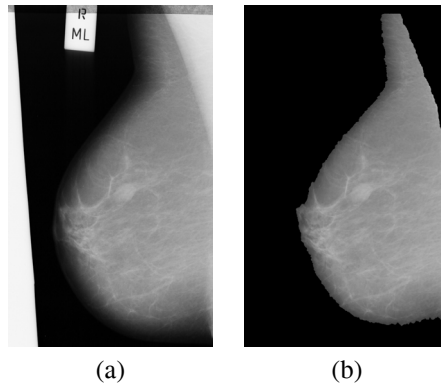


Figure 1. The breast profile segmentation. (a) initial mammogram and, (b) breast area alone.

a portion of the mammogram which is suspicious to be a mass. Therefore, an initial database of RoIs were previously classified as depicting a true mass or not, and the objective of their work was to find which of those previously classified images was closer to the query one. The novelty of this approach is that they used a mutual information metric to compare the query image with the images of the database.

In this paper we propose a novel method for mass segmentation combining the above approaches: starting from a reduced set of templates and using mutual information as the similarity metric, the algorithm will be able to determine if there is a mass on a mammogram. The remainder of this paper is structured as follows: Section 2 describes the proposed segmentation method. Experimental results proving the validity of our proposal are shown in Section 3, and finally, conclusions are given in Section 4.

2. Methodology

As is shown in Figure 1(a), in a mammogram often is possible to distinguish among different regions: the dark background, the pectoral muscle, the breast area and some labels and annotations. Therefore, the first step of our algorithm is to extract the breast area from all the other regions.

2.1. Breast Profile Segmentation

The segmentation of the breast from its surrounding background and pectoral muscle, is based on a previous developed work [13]. Roughly, this method is based on two steps:

- The first step segments the breast area plus the pectoral muscle from the background and annotations. For such task, the algorithm uses, first, an adaptive threshold in order to segment the background, and following, counting the number of pixels of the connected components, the algorithm is able to find the breast plus the muscle as the biggest region.
- The second step segments the pectoral muscle from the breast area. First, a seed is placed automatically inside the pectoral muscle by exploiting the fact that the pectoral muscle is always located in the top corner of the image. Following, an

adaptive region growing algorithm is used to segment the pectoral muscle. The thresholds are found for each mammogram by using its histogram information and assuming that the pectoral muscle is brightness and uniform than the rest of the breast.

Therefore, when this algorithm finishes, only the breast area is selected in the image, as is graphically shown in Figure 1(b).

2.2. Mass Segmentation

Once the breast is segmented, the pattern matching algorithm looks for possible masses. A pattern matching approach starts by defining a template, in this work, a tumour-like template. The definition of the template is based on the approach of Lai et al. [9], who defined the tumour based on three characteristics: brightness contrast, uniform density and circular shape. Figure 2 shows a 5-pixel radius template following such approach. The circular patch of ones in the center represents a tumour area having uniform density. The ring of zeros represents the “don’t care” area to account for some of the shape variability. Finally, the outer edge of the template is filled with minus ones to represent the dark background. Obviously, one of the drawbacks of this algorithm is its poor performance in detecting spiculated masses.

On the other hand, the work of Tourassi et al. [16] is based on the comparison of one RoI among all the others, in order to find if the RoI depicts a (known) true mass. Their work is based on a mutual information approach, were, given two images A and B, the mutual information is expressed as:

$$MI(A, B) = \sum_x \sum_y P_{AB}(x, y) \cdot \log_2 \left(\frac{P_{AB}(x, y)}{P_A(x)P_B(y)} \right) \quad (1)$$

where $P_{AB}(x, y)$ is the joint probability of the two images based on their corresponding pixels values and $P_A(x)$ and $P_B(y)$ are the marginal probabilities of the variables x and y which are the image and the template pixel values. Note that, as we have stated, for using this approach a huge database of known images (i.e., images containing a mass and images not containing a mass) is necessary. Moreover, if the RoI depicts a true mass not present in the database, the algorithm could fail.

The main purpose of our work is, therefore, to take profit of the advantages of both approaches: the use of few general templates to detect masses, and the use of the mutual information in the similarity metric. This way we will be able to compare which metric is better to detect objects with shapes slightly different from the template itself.

As we want to use the similarity metric defined in Equation 1, we need to obtain a compatible template. Note that the template defined by Lai et al. [9] is based on 0 and 1, while Equation 1 is based on gray-level values. This way, we define the template by using the following steps:

1. Calculate the mean of all pixels in the breast.
2. Replace the -1 's of the template by values inferiors to the mean.
3. Replace the 0 's of the template by the mean.
4. Replace the 1 's of the template by values superiors to it.

				-1	-1	-1			
				-1	-1	0	-1	-1	
			-1	-1	0	1	0	-1	-1
		-1	-1	0	1	1	1	0	-1
		-1	0	1	1	1	1	1	0
		-1	-1	0	1	1	1	0	-1
			-1	-1	0	1	0	-1	-1
				-1	-1	0	-1	-1	
					-1	-1	-1		

Figure 2. A tumour-like template for matching with tumours of five pixels in diameter [9].

5. Vary the diameter of the template among different pixels in order to detect masses with different sizes.

In contrast with the work of Lai et al. [9], we consider that the masses have not uniform density. Note that using the above description we let the pixels inside the mass to have whatever value superior to the mean.

3. Experimental Results

With the aim to demonstrate the validity of our approach, we also implemented the original algorithm of Lai et al. [9]. This way, we are able to compare and determine the advantages and constraints of both approaches. We used three different patterns for the original algorithm: one to detect round masses and two for detecting elliptic ones. In contrast, we used only the circular template for the proposed algorithm based on mutual information. Note that we cannot compare our approach with the work of Tourassi et al. [16] because the objective of both works are quite different: while we look for masses in a mammogram, the aim of their work is to retrieval the most similar RoI to a query RoI.

The implemented algorithms have been evaluated using a set of 120 mammograms from the MIAS database [15]. Of this set, 40 mammograms contained masses (which were manually segmented by an expert), whilst the other 80 represented normal cases. The performance of both algorithms have been analysed using ROC and FROC curves [11].

The first experiment was related to the capability of the algorithms to distinguish among mammograms with and without masses, i.e. the capability of the algorithms to detect masses. Therefore, FROC analysis is used in the evaluation. A FROC curve indicates the true positive rate as a function of the number of false positives per image. We consider a true positive, if there are a 50% overlap between the manually segmented and the automatically detected regions. As is shown in Figure 3, although the proposed approach has a tendency to produce a large number of false positives at high sensitivity rates, it is clearly an improvement over the original algorithm, which suffers from a drastically oversegmentation problem.

Once the mammograms containing masses are detected, the second experiment evaluates the accuracy with which the masses have been detected, and therefore, ROC analysis is used. A ROC curve is similar to a FROC one, but with the difference that it indicates the true positive rate as a function of the false positive rate. Moreover, the percentage

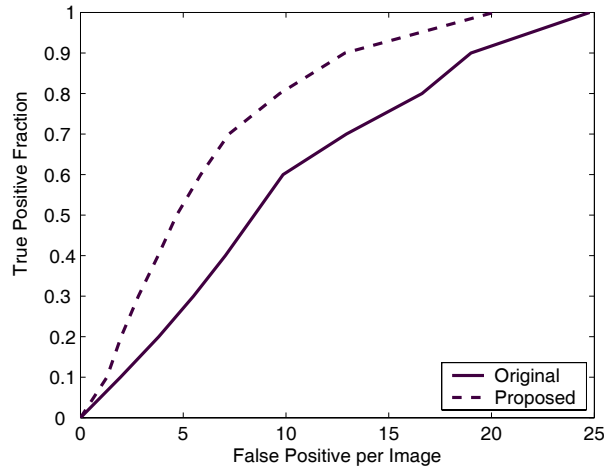


Figure 3. FROC analysis of the algorithm over the set of 120. The continuous line shows the results found by using the original algorithm while the dashed line shows the proposed algorithm. It is clear that the proposal performs better than the original one.

value under the curve (A_z) is an indication for the overall performance of the observer, and is typically used to analyse the performance of the algorithms.

The overall performance over the 40 mammograms of both approaches were: 76.0 ± 7.9 for the original algorithm and 86.1 ± 9.5 for the proposed one. Therefore, we can see that the proposed algorithm increase considerably the performance over the original one. Let us, in what follow, analyse the effect of some parameters of the lesion over both algorithms. In concrete we analyse the behaviour of both algorithms according to the shape and size of the lesions.

3.1. Lesion Shape Influence

The set of annotations of MIAS database lets us to distinguish among two kinds of lesion shape: circular lesions and spiculated lesions. The first ones are related with those lesions that have approximately an elliptic shape, while the latter are related with those lesions which are lesions that usually have a central mass lesions surrounded by an abnormal pattern of radiating linear structures (spicules).

The lesion shape has a strong influence on the performance of both segmentation algorithms, as is shown in Table 1, where the mean and standard deviation values of A_z when segmenting the 40 mammograms with masses are classified according this parameter. Note that the results show that both algorithms segment more accurately circular masses than spiculated ones. This is an obvious result according the shape of the templates used. However, note that using the proposed approach, the influence of the lesion shape is much less important than using the original one.

3.2. Lesion Size Influence

Table 2 shows the mean and the standard deviation of the values of A_z according the size of the masses. We can see that both approaches have better performances when the

Table 1. Influence of the lesion shape for the original algorithm and the proposed one. The results show mean and standard deviation A_z values.

	Circ	Spic
$A_z(\text{Original})$	77.3 ± 8.9	72.2 ± 6.2
$A_z(\text{Proposed})$	84.6 ± 10.8	82.6 ± 7.9

Table 2. Influence of the lesion size (in mm^2) for the original algorithm and the proposed one. The results show mean and the standard deviation A_z values.

	< 2	2 – 4	4 – 6	6 – 8	8 – 15	> 15
$A_z(\text{Or})$	80.2 ± 5.3	78.8 ± 7.6	77.1 ± 8.4	74.7 ± 8.5	73.5 ± 9.5	70.9 ± 5.2
$A_z(\text{Pr})$	93.1 ± 3.4	91.7 ± 9.2	83.0 ± 9.1	82.3 ± 6.6	83.0 ± 5.3	78.3 ± 8.5

mass is small. However, note that the proposed algorithm is much more accurate than the original one, reaching an A_z value of more than 93%, which means that it almost detect their real shape. However, note that this high rate is due to the fact that small masses have rounded shapes and small spicules, if any.

On the other hand, both approaches suffer a drastical decrease of their accuracy when the masses present in the mammograms have larger size. This is due to the fact that such masses can appear in a great variety of shapes, including large spicules, while our templates were still circular.

4. Conclusions and Further Work

In this paper we presented a new approach to the detection of masses in mammographic images. We take as a starting point, an old pattern matching algorithm and, we improve it using a similarity metric based on mutual information. Results are quantitatively evaluated using ROC and FROC analysis, and demonstrate that a mutual information based metric is useful to detect masses with slightly different shapes to the template one. Moreover, we have seen that the effectiveness of the proposed algorithm increases as the size of the masses decreased. Note that this capability is directly related to the detection of early breast cancer, which is demonstrated a crucial issue for improving its posterior treatment. The main drawback of this approach, however, is the large number of false positive it detects. Further work is directed in two directions. First, the inclusion of a postprocessing step in order to reduce the number of false positive. And secondly, the inclusion of a module for detecting spicules in those masses detected with a large size.

Acknowledgements

We sincerely thank Dr. Erika R.E. Denton, of the Dept. of Breast Imaging of the Norfolk and Norwich University Hospital in Norwich (U.K.) for accurately segmenting the masses.

This work was supported by the Spanish commission MCyT (Ministerio de Ciencia y Tecnología) under grant TIC2002-04160-C02-01, FEDER (Fondo Europeo de Desarrollo Regional), and CeRTAP (Centre de Referència en Tècniques Avançades de Producció)

References

- [1] American Cancer Society. Breast cancer: facts and figures. 2003-04. ACS, 2003.
- [2] L. Basset and R. Gold. *Breast Cancer Detection: Mammograms and Other Methods in Breast Imaging*. Grune & Stratton, New York, 1987.
- [3] D. Brzakovic, X.M. Luo, and P. Brzakovic. An approach to automated detection of tumors in mammograms. *IEEE Trans. Med. Imag.*, 9(3):233–241, 1990.
- [4] Eurostat. Health statistics atlas on mortality in the european union. *Official Journal of the European Union*, 2002.
- [5] N. Karssemeijer and G. te Brake. Combining single view features and asymmetry for detection of mass lesions. In *Int. Work. on Dig. Mammography*, pages 95–102, June 1998.
- [6] H. Kobatake, M. Murakami, H. Takeo, and S. Nawano. Computerized detection of malignant tumors on digital mammograms. *IEEE Trans. Med. Imag.*, 18(5):369–378, 1999.
- [7] D. Kopans. *Breast Imaging*. Lippincott-Raven, Philadelphia, 1998.
- [8] M. A. Kupinski and M.L. Giger. Automated seeded lesion segmentation on digital mammograms. *IEEE Trans. Med. Imag.*, 17(4):510–517, 1998.
- [9] S.-M. Lai, X. Li, and W.F. Bischof. On techniques for detecting circumscribed masses in mammograms. *IEEE Trans. Med. Imag.*, 8(4):377–386, 1989.
- [10] L. Li, W. Qian, and L.P. Clarke. Digital mammography: computer assisted diagnosis method for mass detection with multi-orientation and multiresolution wavelet transform. *Acad. Radiol.*, 4(11):724–731, 1997.
- [11] C.E. Metz. Evaluation of digital mammography by roc analysis. In *Int. Work. on Dig. Mammography*, pages 61–68, Chicago, USA, June 1996.
- [12] N. Petrick, H.P. Chan, B. Sahiner, and M.A. Helvie. Combined adaptive enhancement and region-growing segmentation of breast masses on digitized mammograms. *Med. Phys.*, 26(8):1642–1654, 1999.
- [13] D. Raba, A. Oliver, J. Martí, M. Peracaula, and J. Espunya. Breast segmentation with pectoral muscle suppression on digital mammograms. In *Ib. Conf. on Patt. Rec. and Image Anal.*, pages 471–478, Estoril, Portugal, June 2005.
- [14] B. Sahiner, H.P. Chan, D. Wei, N. Petrick, M.A. Helvie, D.D. Adler, and M.M. Goodsit. Image feature selection by a genetic algorithm: Application to classification of mass and normal breast tissue. *Med. Phys.*, 23:1671–1684, 1996.
- [15] J. Suckling, J. Parker, D.R. Dance, S. Astley, I. Hutt, C.R.M. Boggis, I. Ricketts, E. Stamatakis, N. Cerneaz, S.L. Kok, P. Taylor, D. Betal, and J. Savage. The mammographic image analysis society digital mammogram database. In *Int. Work. on Dig. Mammography*, pages 211–221, June 1994.
- [16] G.T. Tourassi, R. Vargas-Vorecek, D.M. Catarious, and C.E. Floyd. Computer-assisted detection of mammographic masses: A template matching scheme based on mutual information. *Med. Phys.*, 30(8):2123–2130, 2003.
- [17] B. Zheng, W.F. Good, D.R. Armfield, C. Cohen, T. Hertzberg, J.H. Sumkin, and D. Gur. Performance change of mammographic cad schemes optimized with most-recent and prior image databases. *Acad. Radiol.*, 10(3):238–288, 2003.
- [18] R. Zwiggelaar, T.C. Parr, J.E. Schumm, I.W. Hutt, C.J. Taylor, S.M. Astley, and C.R.M. Boggis. Model-based detection of spiculated lesions in mammograms. *Med. Image Anal.*, 3(1):39–62, 1999.