Neurocomputing 120 (2013) 61–71

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Automatic segmentation of breast masses using enhanced ICA mixture model

Patricia B. Ribeiro [a],*, Roseli A.F. Romero [b], Patrícia R. Oliveira [c], Homero Schiabel [a], Luciana B. Verçosa [a]

[a] USP/EESC/Department of Electrical Engineering, São Carlos, Brazil
[b] USP/ICMC/Department of Computer Science, São Carlos, Brazil
[c] USP/School of Art, Sciences and Humanities of University of Sao Paulo, São Paulo, Brazil

ARTICLE INFO

Article history:
Received 24 April 2012
Received in revised form
31 July 2012
Accepted 22 August 2012
Available online 27 March 2013

Keywords:
Image segmentation
Mammographic images
Unsupervised learning techniques
Clustering
Visual metrics

ABSTRACT

Image segmentation is considered, among all the stages of image processing, the most critical stage of data processing, because a good classification is dependent on the features extracted from the segmented images. In this work, we are proposing to use the technique called Enhanced ICA Mixture Model (EICAMM) for automatic segmentation of breast masses, aiming to comparing it to other segmentation methods known for segmentation of medical images such as Watershed, Self-Organizing Map (SOM), K-means and Fuzzy C-means techniques. For the analysis of the results, it was used Jaccard similarity measure for comparing the result obtained by the segmentation techniques with that one obtained by an expert opinion. All images considered in this work were segmented and then analyzed by us to improve the segmentation performed by an expert and to detect lesion shape for further classification.

These models have been applied for the segmentation of suspicious masses in digital mammographic images, including images of dense breasts. The obtained results show a good performance of EICAMM that was the unique technique able to detect masses in dense breast interest region in preprocessed images and in original images. In this way, EICAMM could be considered as a good alternative approach to be applied for breast masses classification.

1. Introduction

The X ray mammography is the most common technique used by radiologists in the analysis and diagnosis of breast cancer. In [1], although it can be considered as the most sensitive test for screening of breast cancer, since it reduces mortality rates by up to 25%, its interpretation is very difficult since 10–30% of breast cancer lesions are missed during routine examinations [2–5].

The main lesions identified on mammography are: microcalcifications, which is one of the first signs of tumor formation with a high degree of suspicion of malignancy, and breast masses, which are responsible for the most cases of breast cancer. The lesions are described by their shapes (see Fig. 1) and boundary proprieties (see Fig. 2). A benign tumor has well-defined margins, whereas the cancer is characterized by a indistinctive border that becomes more spiculated with time evolution [6].

Distortions in the interpretation and classification of suspicious lesions by experts involve a larger number of unnecessary biopsies [7–9]. In other words, 65–85% of breast biopsies are performed in benign masses [6,10–13].An important step in Computer-Aided Diagnosis (CAD*x*) schemes in medicine is the image analysis automation. The image segmentation is the first step in such analysis, which is very important stage, because a good classification depends on the vast majority of features extracted from these images. Therefore, a variety of image segmentation techniques has been studied [14–18].

According to Gonzales and Woods [19], segmentation is considered, among all the stages of image processing, the most critical stage of data processing. So, to prevent any errors or distortions that can cause an incorrect diagnostic, in this work, a comparative analysis of different segmentation techniques is performed aiming to indicate the most adequate for the breast masses segmentation.

The accuracy in the representation of nodules is important for the classification of ROI according to its lesion to allow an appropriate feature extraction. The selection of attributes that effectively differentiates the precise standards required in the classification process can prevent errors in the categorization of the structure (as benign or malignant, for example).

It is believed that a probable reason to the application of segmentation techniques that, instead of describing the quality node, distort the final image segmentation, which reduces the data

---

* Corresponding author. Tel.: +55 16 3373 9371; fax: +55 16 3373 9371.
E-mail addresses: patriciabellin@yahoo.com.br, bellin@usp.br (P.B. Ribeiro).

0925-2312/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2012.08.062

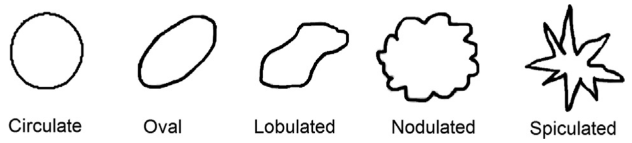

Fig. 1. Typical shapes of breast nodules [6].

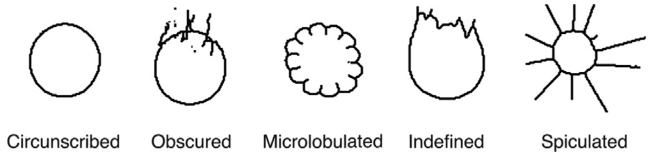

Fig. 2. Types of mass contours [6].

obtained through the extraction of features from the regions of interest.

Artificial Neural Networks (ANN) and Fuzzy Systems can be used as classifiers in CAD*x* schemes [20,21], and their performance depends on several previous stages the process of classification, such as, acquisition of patterns, extraction and selection of features that represent each lesion type [22,23]. In our previous works [24,25], the Self-Organizing Model Enhanced ICA Mixture Model (EICAMM) was proposed as a good alternative for unsupervised classification.

In this paper, we address to investigate the powerful of EICAMM, by the first time, as a tool for digital mammography CAD*x* scheme. In order to analyse its efficiency in relation to other segmentation techniques for breast masses, its performance will be compared with the most common techniques used for pattern classifications such as, watershed [27,28], k-means [29], fuzzy C-means [30] and finally, self organizing map (SOM) [31].

A general scheme of the module CAD*x* breast masses in development in our research group is presented in Fig. 3, to turn clearer what we are addressing in the present paper. In this figure, after the cutting of interest ROI, the following steps must be processed: (a) classification of regions of interest (ROIs) in images with nodule and normal (no findings clinical), (b) segmentation of the IRs, (c) classification according to the contour and (d) classification according to form. The item (b) is focalized in this paper.

To measure the quality of the segmentation obtained by the mentioned different techniques, the metric Jaccard proposed by [28] will be used. All images considered in this work were segmented and then analyzed by us to improve the segmentation performed by an expert and to detect lesion shape for further 1`classification.

2. Approaches to unsupervised image segmentation

An extensive comparative study among different unsupervised techniques that can be applied to segment mammography images has been done. Two approaches for image segmentation have been used. They are clustering-based methods and the watershed segmentation algorithm, which have been widely used in classical image processing applications [24].

Clustering is an essential and frequently used tool in unsupervised pattern recognition and data mining. Its main goal is to divide a set of data into meaningful groups whose members are similar in some way. In general, approaches for clustering represent samples by feature vectors and can be categorized according to the way they mathematically define a cluster: based on a distance measure or based on a probabilistic model [32].

Distance-based clustering algorithms define a suitable distance measure to determine the proximity between two data points (or clusters). Many distance criteria for assessing similarity degrees have been proposed in the literature. Some of the most classical measures are the Euclidean distance (commonly used in SOM, k-means and fuzzy c-means algorithms), the Manhattan distance and the statistical correlation [19].

Mixture model-based (or probabilistic) clustering aims to discover groups of data, which have been generated by different processes. In a typical mixture model, each cluster is represented by a parametrical statistical distribution and the entire data set is modeled by a linear combination (mixture) of such distributions. This probabilistic framework presents the advantage that it addresses the clustering problem taking into account the uncertainty inherent to important issues, such as cluster locations and shapes. In this work, a mixture model formulation based on Independent Component Analysis (ICA) is considered [24,25].

The way how the main techniques have been utilized for processing breast masses is described to follow. For EICAMM method is presented also its respective algorithm, due to fact that it is being applied to the breast mass classification for the first time.

2.1. Automatic segmenters

2.1.1. Enhanced ICA Mixture Model (EICAMM)

ICA Mixture Model (ICAMM) [33] has a goal to overcome a technical limitation of ICA, which relies on the assumption that the sources generating the data are independent. In ICAMM, this assumption was relaxed using the concept of statistical mixture models.

According to Lee, Sejnowski and Lewicki [33], in ICAMM each class is described by a linear combination of non Gaussian independent source densities. The unsupervised classification algorithm estimates the ICA basis matrices and biases vectors (model parameters) for each class using the extended Infomax approach [34], and then applies the Bayes theorem to associate a pattern to the class with the highest posterior probability. The learning rules for the estimation process have been derived by a gradient optimization technique so as to maximize the log-likelihood data function.

Despite promising features of the ICAMM have been reported in [33], some limitations on such approach have been pointed out in Oliveira and Romero [24]. The first one is related to the fact that its learning algorithm is based on a gradient optimization technique and, thus, it can be affected, among other factors, by some drawbacks associated with this kind of approach. Furthermore, Oliveira and Romero [24] showed that some aspects of its formulation and implementation could be improved in order to make the method more efficient, such as incorporating some features of nonlinear optimization methods to the learning procedure. In this sense, the Levenberg–Marquardt technique [35] was integrated to the learning algorithm to ensure and improve the convergence of the model. Finally, an alternative formulation to the generative ICA model has been proposed as well. Such improvements to the ICAMM gave rise to the Enhanced ICA Mixture Model (EICAMM) [24,25].

The procedure for estimating the distribution parameters for each cluster aims to maximize the likelihood function given by:

$$p(X|\theta)=\prod_{t=1}^{T}p(x_t|\theta). \quad (1)$$

where, $X=\{x_1,\ldots,x_T\}$ is the input data, $\theta=\{\theta_1,\ldots,\theta_k\}$ are unknown parameters for each $p(x|C_k,\theta_k)$ where C_k denotes the class k. It is assumed that the number of classes, k, is known. In model EICAMM [25], the input data of each class is given by:

$$x_t=A_K(s_k+b_k). \quad (2)$$