journal homepage: www.intl.elsevierhealth.com/journals/cmpb

A fully automated scheme for mammographic segmentation and classification based on breast density and asymmetry

Stylianos D. Tzikopoulos[a,*], Michael E. Mavroforakis[b], Harris V. Georgiou[a], Nikos Dimitropoulos[c], Sergios Theodoridis[a]

[a] National and Kapodistrian University of Athens, Dept. of Informatics and Telecommunications, Panepistimiopolis, Ilissia, Athens 15784, Greece
[b] University of Houston, Department of Computer Science, 501 P.G. Hoffman Hall, Houston, TX 77204-3010, USA
[c] Delta Digital Imaging, Semitelou 6, Athens 11528, Greece

ARTICLE INFO

Article history:
Received 11 February 2010
Received in revised form
23 November 2010
Accepted 30 November 2010

Keywords:
Automated mammogram segmentation
Breast boundary
Pectoral muscle
Breast density
Breast asymmetry
Nipple detection

ABSTRACT

This paper presents a fully automated segmentation and classification scheme for mammograms, based on breast density estimation and detection of asymmetry. First, image preprocessing and segmentation techniques are applied, including a breast boundary extraction algorithm and an improved version of a pectoral muscle segmentation scheme. Features for breast density categorization are extracted, including a new fractal dimension-related feature, and support vector machines (SVMs) are employed for classification, achieving accuracy of up to 85.7%. Most of these properties are used to extract a new set of statistical features for each breast; the differences among these feature values from the two images of each pair of mammograms are used to detect breast asymmetry, using an one-class SVM classifier, which resulted in a success rate of 84.47%. This composite methodology has been applied to the miniMIAS database, consisting of 322 (MLO) mammograms -including 15 asymmetric pairs of images-, obtained via a (noisy) digitization procedure. The results were evaluated by expert radiologists and are very promising, showing equal or higher success rates compared to other related works, despite the fact that some of them used only selected portions of this specific mammographic database. In contrast, our methodology is applied to the complete miniMIAS database and it exhibits the reliability that is normally required for clinical use in CAD systems.

1. Introduction

Breast cancer, i.e., a malignant tumor developed from breast cells, is considered to be one of the major causes for the increase in mortality among women, especially in developed countries. More specifically, breast cancer is the second most common type of cancer and the fifth most common cause of cancer death according to Nishikawa [1].

While mammography has been proven to be the most effective and reliable method for the early detection of breast cancer, as indicated by Siddiqui et al. [2], the large number of mammograms, generated by population screening, must be interpreted and diagnosed by a relatively small number of radiologists. In addition, when observing a mammographic image, abnormalities are often embedded in and camouflaged by varying densities of breast tissue structures, resulting in high rates of missed breast cancer cases as mentioned by

* *Corresponding author*. Tel.: +30 210 7275104; fax: +30 210 7275214.
E-mail address: stzikop@di.uoa.gr (S.D. Tzikopoulos).
URL: http://www.di.uoa.gr/ stzikop (S.D. Tzikopoulos).

0169-2607/$ – see front matter © 2010 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.cmpb.2010.11.016

Wroblewska et al. [3]. In order to reduce the increasing workload and improve the accuracy of interpreting mammograms, a variety of computer-aided diagnosis (CAD) systems, that perform computerized mammographic analysis have been proposed, as stated by Rangayyan et al. [4]. These systems are usually employed as a second reader, with the final decision regarding the presence of a cancer left to the radiologist. Thus, their role in modern medical practice is considered to be significant and important in the early detection of breast cancer.

All of the CAD systems require, as a first stage, the segmentation of each mammogram into its representative anatomical regions, i.e., the breast border, the pectoral muscle and the nipple, as in the work by Ferrari et al. [5]. The breast border extraction is a necessary and cumbersome step for typical CAD systems, as it must identify the breast region independently of the digitization system, the orientation of the breast in the image and the presence of noise, including imaging artifacts. The goal is to exclude the background from the subsequent processing steps, reducing the image file size without losing anatomic information. It should also have a fast running time and be sufficiently precise, in order to improve the accuracy of the overall CAD system.

The *pectoral muscle* is a high-intensity, approximately triangular region across the upper posterior margin of the image, appearing in all the medio-lateral oblique (MLO) view mammograms and in 30–40% of the cranio-caudal (CC) mammograms, as described by Andolina et al. [6]. Automatic segmentation of the pectoral muscle can be useful in many ways, according to Kwok et al. [7] and Ferrari et al. [8]. One example is the reduction of the false positives in a mass detection procedure, because of the similarity between the pectoral region and the mammographic glandular parenchyma. In addition, the pectoral muscle must be excluded in an automated breast tissue density quantification method. *The location of the nipple* is also of great importance, as it is the only anatomical landmark of the breast, as mentioned by Andolina et al. [6], and can therefore serve as a key point for the whole mammographic image. Most CAD systems use the nipple as a registration point for comparison, when trying to detect possible asymmetry between the two breasts of a patient, according to Yin et al. [9]. These automatic methods can also use the nipple as a starting point for cancer detection, as cancer appears in the glandular/ductal (not the fatty) tissue of the breast, which ends at the nipple and appears as a "cone" to the remaining breast area, as mentioned by Knauerhase et al. [10]. In addition, radiologists pay specific attention to the nipple, when examining a mammogram, according to Chandrasekhar and Attikiouzel [11] and Méndez et al.[12].

Another important characteristic of a mammogram is the breast *parenchymal density* with regard to the prevalence of fibroglandular tissue in the breast as it appears on a mammogram. The relation between mammographic parenchymal density levels and high risk of breast cancer was first shown by Wolfe [13], using four distinct classes for breast parenchymal density categorization, leading later to the BI-RADS classification scheme proposed by De Orsi et al. [14] from the American College of Radiology (ACR). Thus, mammographic images with high breast density value should be examined more carefully by radiologists, for both physiological and imaging risk factors, creating a need for automatic breast parenchymal density estimation algorithms. In Masek [15], such algorithms are presented and a new technique, introducing a histogram distance metric, achieves good results. Some existing algorithms, e.g., Bosch et al. [16] and Oliver et al. [17], use the texture information of mammograms, in order to extract more features for breast density estimation.

Radiologists try also to detect possible *asymmetry* between the left and the right breast in a pair of mammograms, as it can provide clues about the presence of early signs of tumors such as parenchymal distortion, according to Homer [18]. Many CAD systems analyze automatically the images of a mammogram pair and provide results for the detection of asymmetric abnormalities by applying some type of alignment and direct comparison, as implemented by Yin et al. [9]. In the works of Ferrari et al. [19] and Rangayyan et al. [20], directional analysis methods are proposed, using Gabor wavelets, in order to detect possible asymmetry.

In this work, we propose a fully automated and complete segmentation methodology as the first stage of a multi-stage processing procedure for mammographic images. Specifically, we have chosen to implement and apply the algorithm presented by Masek [15] for breast boundary extraction, as the first step of the composite processing procedure; for the second step of pectoral muscle estimation, we enhanced the algorithm presented by Kwok et al. [7] in order to achieve improved results; as a third step, we propose a new nipple detection technique, using the output of the breast boundary extraction procedure, when the nipple is in profile; that is, when it is projected on the background area of the mammogram, which is the recommended and usual case. The last algorithm, that is proposed in this work, besides locating the nipple point, can also serve as an improvement for the existing breast boundary algorithm, which misses the nipple if it is in profile. The improvement is obtained when updating the breast boundary, in order to include the detected nipple. Furthermore, as a fourth step, a new breast parenchymal density estimation algorithm is proposed, using segmentation, first-order statistics and fractal-based analysis of the mammographic image for the extraction of new statistical features, while the classification task is performed using support vector machines (SVMs). Finally, a new algorithm is proposed for breast asymmetry detection, using the feature values already extracted from the breast parenchymal density estimation step, using an one-class SVM classifier. Both techniques achieve high success rates, often higher than the corresponding values of other algorithms in the relevant literature, while simpler and faster feature extraction methods have been employed. Our methodology has been tested on all the 322 mediolateral oblique view mammograms of the complete miniMIAS database, which is provided by Suckling et al. [21], giving prominent results according to specific statistical measures and evaluation by expert radiologists, even in the case of such a difficult (very noisy) mammographic dataset.

The rest of this paper is organized as follows: In Section 2, the mammographic image database used is described. The pre-processing techniques, the segmentation algorithms, the breast parenchymal density estimation method and the asymmetry detection scheme are described in Section 3. Section 4 presents the results obtained by the proposed algorithms,

while the discussion and the conclusions are summarized in Section 5.

2. Dataset

The methodology presented in this work was applied on the complete miniMIAS database [21]. It is available online freely for scientific purposes and consists of 161 pairs of mediolateral oblique view mammograms. The images of the database originate from a film-screen mammographic imaging process in the United Kingdom National Breast Screening Program. The films were digitized and the corresponding images were annotated according to their breast density by expert radiologists, using three distinct classes: Fatty (F) (106 images), Fatty-Glandular (G) (104 images) and Dense-Glandular (D) (112 images), similar to Mavroforakis et al. [22]. Any abnormalities were also detected and described, including calcifications, well-defined, spiculated or ill-defined masses, architectural distortion or asymmetry. Each pair of images in the database is annotated as Symmetric (146 pairs) or Asymmetric (15 pairs). The severity of each abnormality is also provided, i.e., benignancy or malignancy.

Typical mammographic images are shown in Fig. 1. The presence of high levels of noise and imaging artifacts is readily observed; this makes the segmentation of the image a demanding task. Moreover, speckle noise was added through the original digitization processing of the film mammograms. The original 0.2 mm/pixel images were resized to 0.4 mm/pixel, as in Kwok et al. [7] and Chandrasekhar and Attikiouzel [11], in order to reduce the required computational time, whereas the initial bit depth of 8 bits was preserved thoughout all the experiments and processing steps. It should be noted that the very high noise levels introduced in the digital images makes the miniMIAS dataset a very hard testbed for our methodology and this is a major reason of adopting it.

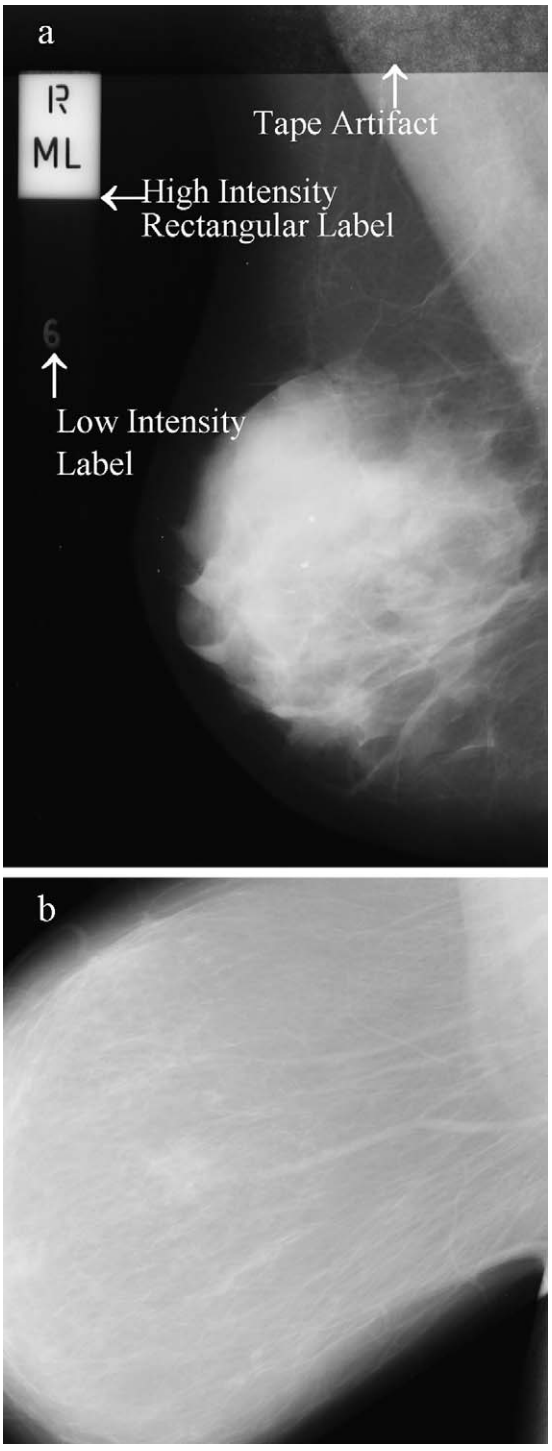

Fig. 1 – Images of the database used: (a) types of noise observed at a mammogram and (b) an example of a mammogram with the breast cut off.

3. Methodology

3.1. *Mammogram image pre-processing*

Image pre-processing techniques are necessary, in order to find the orientation of the mammogram, remove the noise and enhance the quality of the image. Thus, (i) an algorithm to deduce the orientation of the image is implemented, (ii) the noise is estimated according to a specific scheme and (iii) an image filtering technique is adopted for enhancement.

3.1.1. *Image orientation*

The orientation of the mammogram is determined according to Masek [23]. The image is rotated and reflected, so that the chest wall location, i.e., the side of the image containing the pectoral muscle, is on the left side of the image and the pectoral muscle is at the upper-left corner of the image. An example is shown in Fig. 2. In Fig. 2a, the initial image is shown, on which the algorithm is applied. In order to determine the chest wall location, the decreasing pixel intensity of the breast tissue near the skin-air interface (breast boundary) is used, as Fig. 2b displays. This tissue is located by employing the minimum cross-entropy thresholding technique, proposed by Brink and Pendock [24], twice in the original image. By estimating the first derivatives in these pixel transition areas, using the appropriate convolution masks, we can determine the chest wall location. The image is rotated, in order for the chest wall location to be placed on the left side of the image, resulting to the image of Fig. 2c. Next, the top of the image

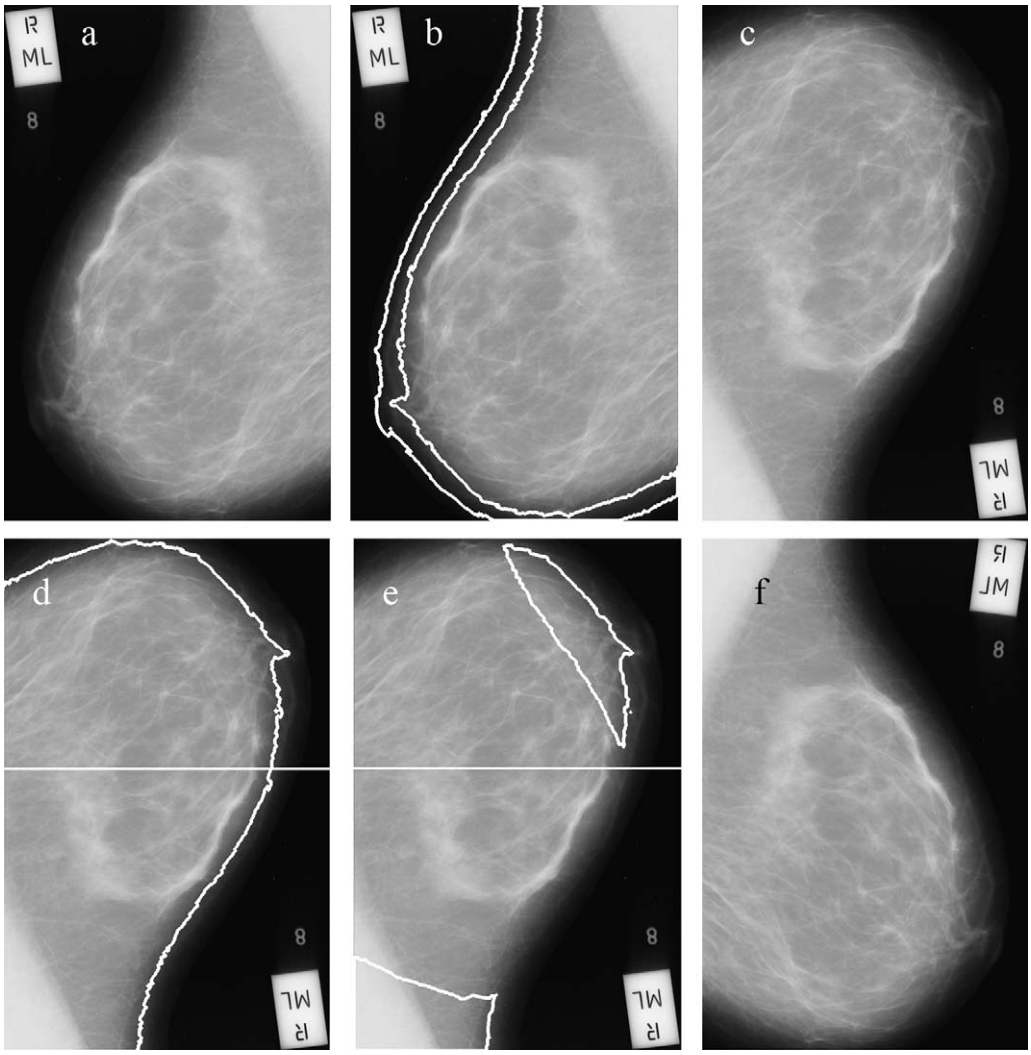

Fig. 2 – The different steps of the image orientation procedure: (a) initial image, (b) the pixels of the breast tissue near the skin-air interface, (c) initial image rotated, (d) the vertical centroid extracted, (e) the asymmetric regions, (f) the final reflected image.

is determined: At first, we extract the vertical centroid of the image, as the row dividing the skin tissue mask into two equal parts, as Fig. 2d shows. Then, the asymmetric regions with respect to the vertical centroid are estimated (Fig. 2e). We assert that the asymmetric region closest to the right side of the vertical centroid is the tip of the breast. The image is flipped vertically, if needed, to place this asymmetric region below the vertical centroid, resulting in an image the right way up as in Fig. 2f.

3.1.2. *Noise estimation*

As in typical film scanned mammographic images, in the images of miniMIAS database several types of noise and imaging artifacts are present, as Fig. 1a shows. Our methodology estimates those regions and excludes them from the remaining process.

Noise corresponding to high values of optical densities is refered to as "high intensity noise". Examples are the labels or the scanning artifacts of Fig. 1a. In order to detect these regions, an existing algorithm, that uses a combination of the 2-level minimum cross entropy thresholding technique [24] as well as of logical and morphological operations, is implemented.

In Fig. 1a, we can also observe "tape" artifacts. These are defined as markings left by tapes or other shadows presenting themselves as horizontally running strips. This horizontal line, corresponding to their edges, is used for the segmentation of this type of "noise" as in [15]. The methodology first detects the high intensity noise and determines the orientation of the image. Then the image is rotated and flipped, so as to enclose the pectoral muscle on the upper left corner, according to the procedure described in the previous subsection. Then, the horizontal edges of the image are enhanced, using a 3×3 horizontal Sobel mask, described in detail by Sobel [25]. The tape artifact detection is completed by adopting the Radon transform proposed by Radon [26] and performing it on the left-half of the edge-enhanced image containing the pectoral muscle. Obviously, the rotation angle theta of the Radon trans-

form is set to (π/2), in order to compute the projection of the image onto the y-axis.

The already mentioned noise removal techniques are adequate for the separation of the human tissue region from the image background. Other types of noise (besides of the speckle noise, which is discussed separately in the next subsection), such as low intensity noise, are not considered, as their contribution is negligible to the context of this work.

3.1.3. *Image enhancement*

Due to the digitization process, the images of the database contain also speckle noise. In order to enhance the quality of the image and achieve better resulting image quality and, hence, better boundary detection and segmentation results, this type of noise should be eliminated. After trying specific image filtering techniques, the method that was selected to remove this type of noise, preserve the edges of the image and achieve the best segmentation results, is the median filtering, as described by Gonzalez and Woods [27]. The median filter is calculated over a neighborhood of 3×3 pixels.

3.2. *Mammogram segmentation*

3.2.1. *Breast boundary detection*

The adopted method to detect the breast boundary of each mammogram is described in detail in [15]. In this algorithm, two interfaces are estimated and then combined in order to obtain the final one: the row-wise interface, which estimates one pixel from each row as a boundary pixel and the column-wise interface, which also estimates one pixel from each column as a boundary pixel. Each one of the two interfaces is divided into two parts, as shown in Fig. 3a and b, resulting, at the end, to four estimates to be combined in order to obtain the final one (Fig. 3c). Each of them is transformed into a function having one value for each row or column.

The algorithm relies on the idea that the skin-air interface is the smoothest section of identical pixels near the breast boundary. Based on that, we segment the image using a specific threshold, extract the interface and fit polynomials of degree 5 to 10, in order to extract each one of the above four interfaces. Then the square error between the fitted curve and the interface is calculated, "punishing" high values of intensities, in order not to detect contours internal to the breast. This procedure is repeated for several values of the threshold and the final estimate is chosen as the one that results in the minimum error, when compared with the inherently smooth polynomial.

3.2.2. *Pectoral muscle detection*

The region of the pectoral muscle of a mammogram is presented magnified at Fig. 4. In order to detect this region in detail, we used the technique described by Kwok et al. [7], which adopts a two-step segmentation scheme. The first step is called straight line estimation and validation and approximates the boundary as a straight line, as Fig. 4a shows. This line is given as input to the second step of processing, named iterative cliff detection. This procedure iteratively refines the straight line to a curve that depicts the pectoral margin more accurately (Fig. 4b).

At the end of this process, if the bottom end of the estimate is not aligned with the left edge of the image, Region Enclosing is performed. According to this technique, the bottom end is extended by a straight line parallel to the initial straight line estimation. In order to use the updated estimate of the pectoral muscle and not the initial straight line estimation, we extend the bottom end – if needed – by a straight line parallel to the straight line, which best fits the iteratively refined estimate. Using this improvement of the existing algorithm, we achieve better results, as it is obvious from Fig. 4 and analytically presented in subsection 4.1. The initial straight

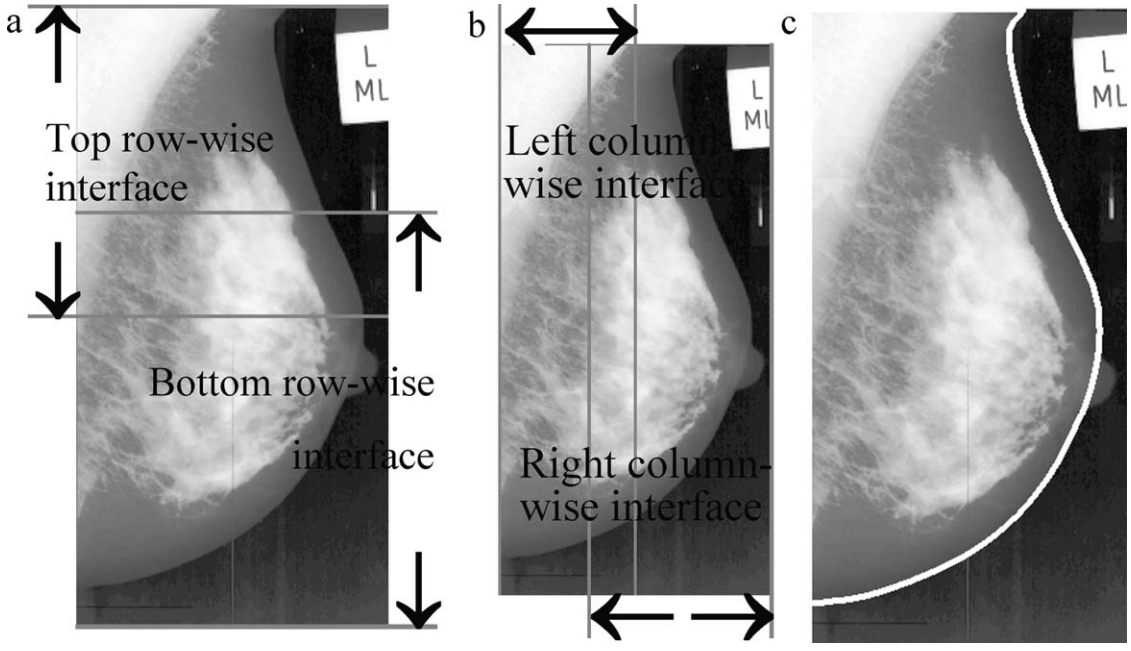

Fig. 3 – The row-wise and the column-wise interfaces estimated (a and b) and combined to determine the final one (c). (a) Row-wise interfaces, (b) column-wise interfaces, (c) final interface.

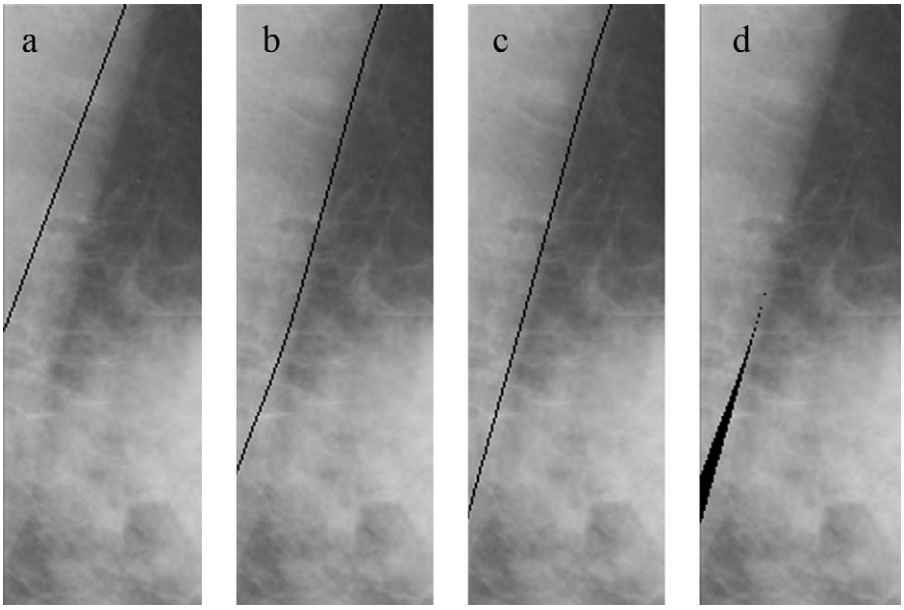

Fig. 4 – Pectoral muscle segmentation procedure: (a) straight line estimation, (b) the final estimate of the algorithm of the bibliography, (c) improvement we propose, (d) difference of the pixels of the two techniques.

line estimation, which is used for the Region Enclosing procedure is not the best one. Thus, the bottom end of the final estimate of Fig. 4b is not aligned with the actual boundary. In Fig. 4c, the line used for the Region Enclosing is not the initial one, but the straight line that best fits the iteratively refined estimate, which results to a better estimate. The difference of the pixels of the two techniques is observed in Fig. 4d.

3.2.3. *Nipple detection*

It is evident from Fig. 3c that the nipple boundary is characterized by high curvature or corner lines. This is the main reason for the inadequacy of the breast boundary estimation algorithm to detect the nipple, when it is in profile. We propose a new technique to detect the nipple whenever this is in profile, using the already estimated boundary.

The regions of a mammogram, which may contain the nipple, correspond to the right-column, bottom-row and top-row interfaces of the breast boundary detection algorithm, as Fig. 3a and b shows. The algorithm uses the thresholds selected for these interfaces. Considering a threshold value, we assume a search area of 10 mm width, which is located on the right of the already detected breast boundary (of a mammogram facing right, Fig. 5a) and we threshold the search area, after performing a 3×3 gaussian filter of 0.5 standard deviation in order to minimize the noise of the background pixels. For each row of the search area, the first zero pixel (the pixel whose value in the initial image is smaller or equal with the threshold value) is detected and all the previous columns are given the value 1, creating a new binary mask S_T, where T is the threshold value. S_T is assumed to be an area that may contain a nipple. The previous procedure is repeated for the minimum and maximum values of the thresholds, as well as for the intermediate values, resulting to several binary masks, some of which are shown in Fig. 5b and c.

Considering a binary mask S_T, an ellipse with moving center at each pixel of the boundary and with variable semi-major and semi-minor axis from 2 mm to 10 mm is drawn, trying to model the possible presence of the nipple, as Fig. 5d shows an ellipse having a 10 mm semi-major axis and 4 mm semi-minor axis, which tries without success to model the nipple. The major axis is considered to be the tangent of the boundary at the specific point. Note that the smallest value of the axis is smaller than the one in Chandrasekhar and Attikiouzel [11], in order to be able to detect smaller nipples. If the pixels of the ellipse, which are located on the right of the boundary have also non-zero values in the binary image S_T, then a possible nipple is detected and those pixels are considered as a region of interest (*nippleROI*).

Subsequently, we use the area $S_{T_{max}}$, defined as the binary mask obtained by the largest value of the thresholds, in order to avoid detecting possible noise pixels as being the nipple. The basic idea is that the segmented mask, which is obtained by the largest value of threshold T_{max} should contain at least one pixel of the nipple; otherwise, we have detected noise as possible nipple area. Thus, a logical AND operator is performed between each region of interest *nippleROI* and $S_{T_{max}}$ and the corresponding *nippleRoi* is discarded if the result is a black binary image not containing any white pixels.

By repeating the previous procedure for all the binary images S_T, we obtain several *nippleROI*'s and we consider the largest of them as the possible nipple, as Fig. 5e indicates.

3.3. *Mammogram classification*

3.3.1. *Breast density estimation*

After the implementation of the complete segmentation scheme, which was previously presented, we adopt a new image pre-processing stage, in order to improve the overall

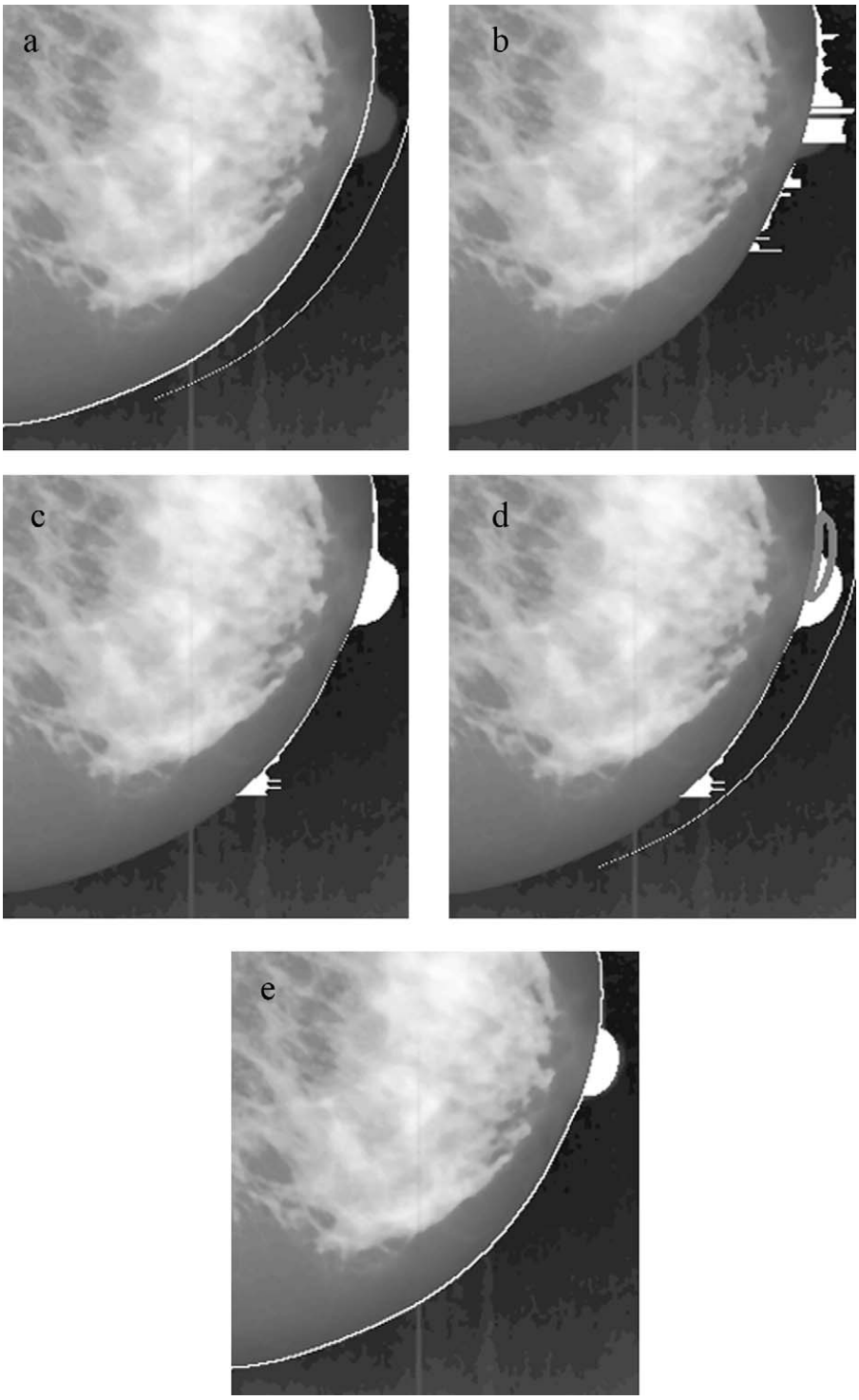

Fig. 5 – Nipple detection procedure: (a) defined search area for a nipple, (b) S_2, (c) S_3 binary masks searched for containing a nipple, (d) an ellipse trying to model the nipple, (e) final nipple estimate.

quality of the images of the database for the breast density classification. This stage includes:

- a gaussian smoothing filter, as described by Gonzalez and Woods [27], with variable kernel size *hsize* and standard deviation *sigma*, in order to remove the noise of the image
- an unsharp filter, as declared by Gonzalez and Woods [27], with custom convolution mask

$$h_{UNSHARP} = \frac{1}{1+a} \cdot \begin{bmatrix} -a & a-1 & -a \\ a-1 & a+5 & a-1 \\ -a & a-1 & -a \end{bmatrix}$$

of variable parameter *alpha*, in order to enhance the edges inside the image

The previous parameters were automatically tuned according to an experimentation scheme. Specifically, the following values were given to the variables and, for each combination

of values, the success rate of the breast density estimation technique was recorded:

- *hsize*: 3×3, 5×5, 7×7, 9×9, 11×11 (pixels $\times$ pixels)
- *sigma*: 0.1, 0.4, 0.7, 1.0
- *alpha*: 0.1, 0.4, 0.7, 1.0

The values that achieved the best success rate were: $hsize = 7 \times 7$, $sigma = 0.4$ and $alpha = 0.7$; these optimal values were used as the baseline for enhancing all the images in the database prior to any breast segmentation and parenchymal analysis.

For the estimation of the features used for the breast density classification scheme, we start from the complete segmentation technique described above. According to this process, each mammogram is analyzed to the following regions, as Fig. 6 shows:

- The initial *I* image (Fig. 6a).
- The background area, labels and artifacts have been excluded, to obtain the *Back* area (Fig. 6b).
- The human-tissue *HuT* area (Fig. 6c), which has been obtained after extracting background, labels, artifacts and noise from the initial image.
- The segmented breast tissue *BrT* area (Fig. 6d), which has been obtained after extracting the pectoral muscle from the human-tissue *HuT* area.

The first two features, used for breast density estimation, are extracted from the *Back* area (no tissue or artifacts). They analyze and model the overall noise levels of the image by estimating the mean and variance of the pixel intensity values of this specific area, as Eqs. (1) and (2) show:

$$F_1 = \mu_{Back} = \frac{\sum_{(i,j) \in Back} I(i,j)}{N(Back)} \tag{1}$$

$$F_2 = \sigma^2_{Back} = \frac{\sum_{(i,j) \in Back} (I(i,j) - \mu_{Back})^2}{N(Back)} \tag{2}$$

where $N(R)$ is the number of pixels in region R.

The features F_3 and F_4 are estimated from the breast tissue (*BrT*) area, according to Eqs. (3) and (4):

$$F_3 = \frac{S_{BrT}}{N(BrT)} \tag{3}$$

$$F_4 = \frac{P_{BrT}}{\mu^2_{BrT}} \tag{4}$$

where S_{BrT} is the graylevel-sensitive surface and P_{BrT} the power of the *BrT* area (Eqs. (5) and (6)):

$$S_{BrT} = \sum_{(x,y) \in BrT} \Big(I(x,y) + 1 + \big|I(x+1,y) - I(x,y)\big| + \big|I(x,y+1) - I(x,y)\big| \Big) \tag{5}$$

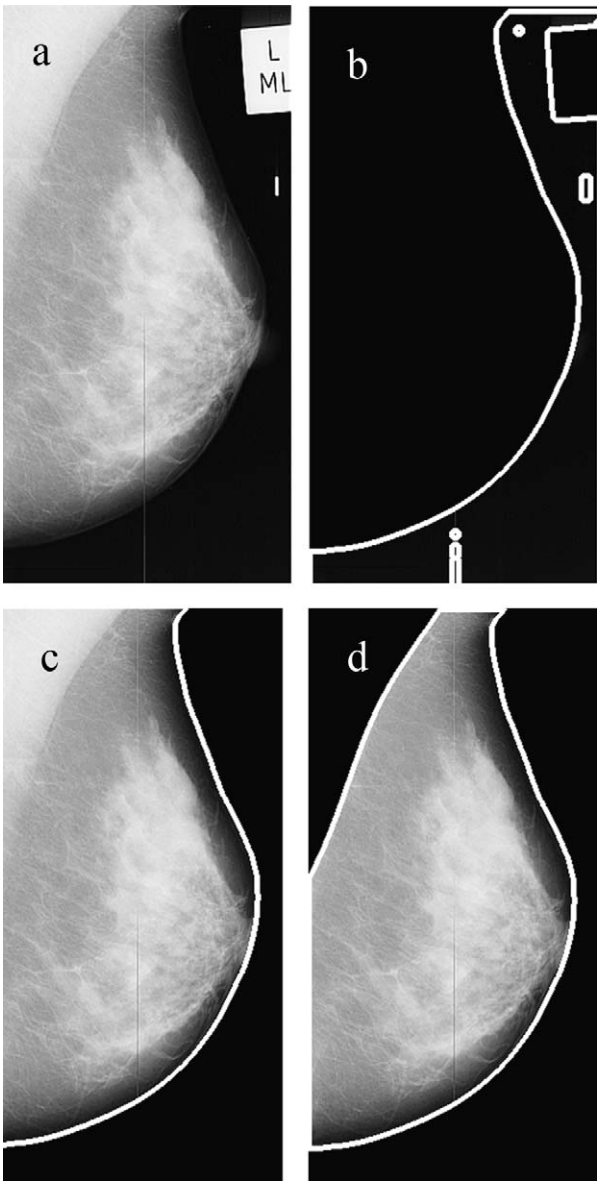

Fig. 6 – Different regions for the feature extraction of the breast density classification: (a) initial image *I*, (b) background *Back*, (c) tissue-rich area *HuT* and (d) breast tissue area *BrT*.

$$P_{BrT} = \sum_{(x,y) \in BrT} \big|I(x,y)\big|^2 \tag{6}$$

Next, an algorithm based on the power spectrum is employed for the computation of a fractal-related feature, as described in Refs. [27,28]. The initial image is resized from 0.4 mm/pixel to the lower resolution of 1.6 mm/pixel (Fig. 7a), after placing black (zero-valued) pixels to the non-*HuT* area. The absolute values of the Discrete Fourier Transform (DFT) of the derived image are estimated and averaged over the four 2-D spectrum quarters. The estimated image is cropped to become square and the logarithmic values over the main diagonal of the spectral image are extracted (Fig. 7b). An exponential function $f(x) = A \exp(Bx) + C$ is fitted to the extracted

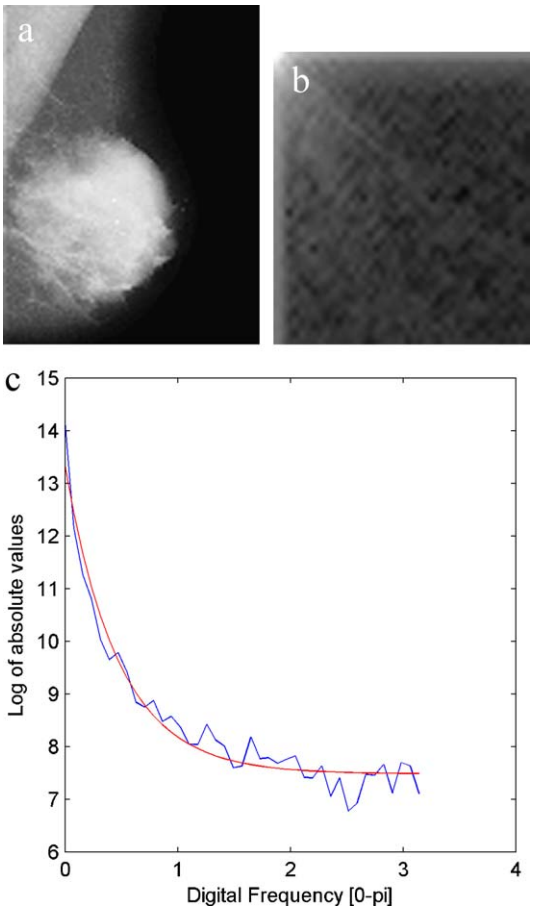

Fig. 7 – The estimation of the fractal-related feature: (a) Initial image resized to lower resolution, (b) logarithmic values of the cropped image of the absolute values of the Fourier transform, (c) fitting an exponential function to the data.

1-D data as Fig. 7c shows and the feature $F_5 = B$ is obtained, as the feature related to the fractal exponent of the texture of the human tissue according to Kaplan [29].

Next, an inner-breast segmentation technique is performed, in order to detect the fibroglandular tissue and its proportion to the whole breast area. For this procedure, the human tissue area *HuT* is used to perform the minimum cross entropy (MCE) thresholding, provided by Brink and Pendock [24], three times, according to the following scheme:

- T is the (baseline) threshold derived from MCE at gray level range $[1, 2^q - 1]$
- T_1 is the threshold derived from MCE at gray level range $[T + 1, 2^q - 1]$
- T_2 is the threshold derived from MCE at gray level range $[T_1 + 1, 2^q - 1]$, where q is the current graylevel depth ($q = 8$)

The value of the threshold T_2 is used to segment the main core of the glandular tissue from the remaining breast area, as Fig. 8b shows. The lower threshold T_1 results to a larger, more detailed description of the glandular tissue, as observed at Fig. 8c. Note that all the possible regions combining the two thresholds T_1 and T_2 are extracted, as Fig. 8a shows. This is due to the importance of the remaining fatty tissue after each segmentation (corresponding to the two thresholds), with regard to shape and size information of the glandular tissue compared to the remaining breast area. Thus, we extract the following regions:

- ROI_1: the pixels $I(x, y)$ with $0 \leq I(x, y) \leq T_2$.
- ROI_2: the pixels $I(x, y)$ with $T_2 < I(x, y) \leq 2^q - 1$.
- ROI_3: the pixels $I(x, y)$ with $0 \leq I(x, y) \leq T_1$.
- ROI_4: the pixels $I(x, y)$ with $T_1 < I(x, y) \leq 2^q - 1$.
- ROI_5: the pixels $I(x, y)$ with $T_1 < I(x, y) \leq T_2$.

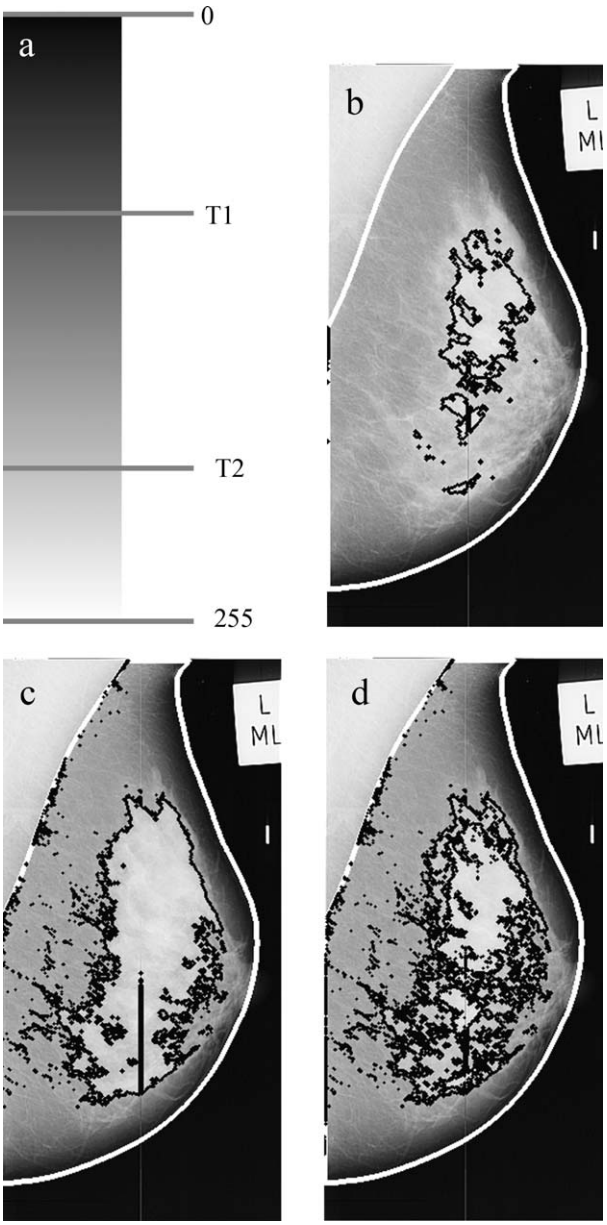

Fig. 8 – Inner-breast segmentation scheme: (a) threshold selection, (b) ROI_1 and ROI_2, (c) ROI_3 and ROI_4 and (d) ROI_5.

Table 1 – Features used for breast density estimation.

$F_1 = \mu_{Back}$	$F_8 = \mu_{ROI_2}$	$F_{15} = \sigma^2_{ROI_4}$
$F_2 = \sigma^2_{Back}$	$F_9 = \sigma^2_{ROI_2}$	$F_{16} = r_4$
$F_3 = \frac{S_{BrT}}{N(BrT)}$	$F_{10} = r_2$	$F_{17} = wr_4$
$F_4 = \frac{P_{BrT}}{\mu^2_{BrT}}$	$F_{11} = wr_2$	$F_{18} = \mu_{ROI_5}$
$F_5 = FE(HuT)$	$F_{12} = \mu_{ROI_3}$	$F_{19} = \sigma^2_{ROI_5}$
$F_6 = \mu_{ROI_1}$	$F_{13} = \sigma^2_{ROI_3}$	$F_{20} = r_5$
$F_7 = \sigma^2_{ROI_1}$	$F_{14} = \mu_{ROI_4}$	$F_{21} = wr_5$

Finally, for each one of the above regions ROI_i, the mean μ_{ROI_i} and the variance $\sigma^2_{ROI_i}$ of the pixel intensities are estimated, according to Eqs. (1) and (2); for the regions ROI_2, ROI_4 and ROI_5 a set of features are also estimated using Eqs. (7) and (8):

$$r_i = \frac{N(ROI_i)}{N(BrT)} \tag{7}$$

$$wr_i = \frac{\sum_{(x,y)\in ROI_i} I(x,y)}{\sum_{(x,y)\in BrT} I(x,y)} \tag{8}$$

where BrT is the segmented breast tissue referred above. This results to a total number of 21 features, as Table 1 shows. For the classification of the images according to the breast density, Classification and Regression Trees (CARTs) as described by Breiman [30] are used. The main motivation for adopting this base classifier was the simplicity of these decision trees. We used three CARTs, equal to the number of the classes. The CART Tr_i is trained to output the value 1 for the images of class i and the value 0 for all the remaining images. Thus, we use an unknown pattern as input to all the CARTs and classify to class j, so that $output(Tr_j) = \max_{1\le k\le 3}\{output(Tr_k)\}$, according to the "one-against-all" classification scheme as described by Theodoridis and Koutroumbas [31]. Another classifier used is the k nearest neighbor classifier, as described by Theodoridis and Koutroumbas [31], whose results are compared with the previous one.

Besides the CARTs and k-nn, the SVM classifier, as presented algorithmically by Mavroforakis and Theodoridis [32] and Mavroforakis [33], was used, in order to classify all the images to the three breast density classes. This classifier maps the data to a high-dimensional space, where the training data are expected to be linearly separable with high probability, and the goal is to design an optimal hyperplane that separates them so that the margin between classes is maximized. SVMs present attractive advantages, such as the uniqueness and sparseness of the solution, and have therefore been successfully applied to a number of applications in various fields, as described by Byun and Lee [34] and Mavroforakis [35], including medical diagnosis, face detection and signal processing. For the SVM classification task, the radial basis function (RBF) kernel was selected. The one-against-one approach was adopted in order to deal with a 3-class problem, using the two class SVM classifier, as described by [31]. For choice of the hyperparameters σ^2 (for the RBF) and the C constant associated with the terms in the SVM's loss function, a grid search tecnhique was adopted.

In order to evaluate the proposed procedure, the leave-one-out methodology was implemented, as described by Theodoridis and Koutroumbas [31]. Accordingly, each one pattern is selected as the unknown one and extracted from the data, resulting to the training set. The classifier is trained and then tested for the unknown pattern. The previous procedure is repeated for all the available data, obtaining the classification results. Apart from the leave-one-out methodology, the leave-one-woman-out algorithm is also used for the evaluation of the system, as presented by Bosch et al. [16]. According to this technique, we leave the two images (left and right breasts) from the same woman out of the training set and use them as the testing set, based on the assumption of the similar tissue features of the both breasts of one woman.

For the sake of reproducibility of the results we mention the optimal values of the parameters $\gamma = (1/2\sigma^2)$ and C, associated with the SVM classifier. Using the leave-one-out evaluation technique we selected $\gamma = 2^{-3}$ and $C = 8$ for the automatic segmentation and $\gamma = 2^{-2}$ and $C = 10$ for the manual segmentation method correspondingly. Using the leave-one-woman-out evaluation technique the values are $\gamma = 2^{-3}$ and $C = 8$ for the automatic segmentation and $\gamma = 2^{-6}$ and $C = 16$ for the manual segmentation method.

3.3.2. *Asymmetry detection*

The basic idea in the feature extraction phase is to use the inner segmentation of the breast, already obtained from the mammographic breast density estimation steps, to provide the necessary means for detecting possible asymmetry between a pair of mammograms. For each mammogram, the features described in Table 2 are calculated. Note that:

- For each one of the regions ROI_2, ROI_4, ROI_5, consider the pixels in ROI_i as 'on' pixels. In order to find the x-axis cummulative projection in the form of a histogram, estimate the number (sum) of 'on' pixels in every row of the image. In the same way, we obtain the y-axis histogram (cummulative

Table 2 – Features used for asymmetry detection.

$F_1 = F^{BRD}_{10}$	$F_{14} = \mu^{Y-AXIS}_{ROI_2}$	$F_{27} = ku^{Y-AXIS}_{ROI_4}$
$F_2 = F^{BRD}_{11}$	$F_{15} = \sigma^{Y-AXIS}_{ROI_2}$	$F_{28} = m^{Y-AXIS}_{ROI_4}$
$F_3 = F^{BRD}_{16}$	$F_{16} = sk^{Y-AXIS}_{ROI_2}$	$F_{29} = \mu^{X-AXIS}_{ROI_5}$
$F_4 = F^{BRD}_{17}$	$F_{17} = ku^{Y-AXIS}_{ROI_2}$	$F_{30} = \sigma^{X-AXIS}_{ROI_5}$
$F_5 = F^{BRD}_{20}$	$F_{18} = m^{Y-AXIS}_{ROI_2}$	$F_{31} = sk^{X-AXIS}_{ROI_5}$
$F_6 = F^{BRD}_{21}$	$F_{19} = \mu^{X-AXIS}_{ROI_4}$	$F_{32} = ku^{X-AXIS}_{ROI_5}$
$F_7 = F^{BRD}_{5}$	$F_{20} = \sigma^{X-AXIS}_{ROI_4}$	$F_{33} = m^{X-AXIS}_{ROI_5}$
$F_8 = N(BrT)$	$F_{21} = sk^{X-AXIS}_{ROI_4}$	$F_{34} = \mu^{Y-AXIS}_{ROI_5}$
$F_9 = \mu^{X-AXIS}_{ROI_2}$	$F_{22} = ku^{X-AXIS}_{ROI_4}$	$F_{35} = \sigma^{Y-AXIS}_{ROI_5}$
$F_{10} = \sigma^{X-AXIS}_{ROI_2}$	$F_{23} = m^{X-AXIS}_{ROI_4}$	$F_{36} = sk^{Y-AXIS}_{ROI_5}$
$F_{11} = sk^{X-AXIS}_{ROI_2}$	$F_{24} = \mu^{Y-AXIS}_{ROI_4}$	$F_{37} = ku^{Y-AXIS}_{ROI_5}$
$F_{12} = ku^{X-AXIS}_{ROI_2}$	$F_{25} = \sigma^{Y-AXIS}_{ROI_4}$	$F_{38} = m^{Y-AXIS}_{ROI_5}$
$F_{13} = m^{X-AXIS}_{ROI_2}$	$F_{26} = sk^{Y-AXIS}_{ROI_4}$	

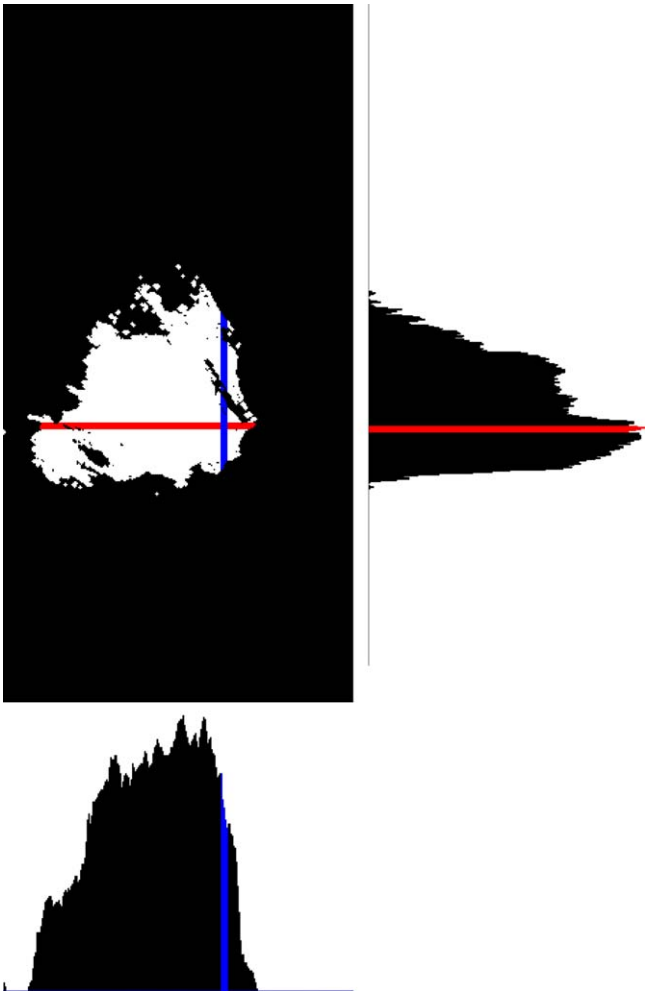

Fig. 9 – The segmentation mask and the *x*-axis (red) and *y*-axis (blue) generated histograms. (For interpretation of the references to color in text, the reader is referred to the web version of the article.)

projection), as shown in Fig. 9. Subsequently, estimate the first-order statistics for each of these histograms, meaning the mean value μ, the standard deviation σ, the skewness sk, the kurtosis ku and the median m.

- The value F_i^{BRD} corresponds to the feature i of the mammographic breast density estimation step (Table 1).

The feature vector of length $N = 38$, described in Table 2, is estimated for each mammogram. However, in our case, we are interested in detecting asymmetry between a pair of mammograms. Thus, we should detect the cases where the values corresponding to the left and the right mammograms differ significantly. Suppose that for the left breast mammogram we have estimated the feature vector f and for the corresponding right breast mammogram the feature vector g. Then, we construct the following differential features of Eqs. (9)–(11), that can be used to detect possible asymmetry between a pair of mammographic images:

$$F_{1-38}^{ASYMMD} = \frac{|f_i - g_i|}{\max(f_i, g_i)} \quad (9)$$

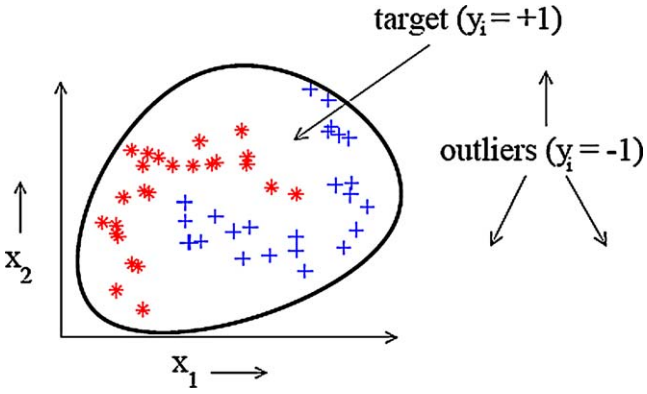

Fig. 10 – One-class classification at a 2-dimensional feature space (x_1, x_2). The classifier is trained according to the target patterns ($y_i = +1$); everything outside is considered as an outlier ($y_i = -1$).

$$F_{39-76}^{ASYMMD} = |f_i - g_i| \quad (10)$$

$$F_{77-114}^{ASYMMD} = |f_i - g_i|^3 \quad (11)$$

where $1 \le i \le 38$, resulting to a feature space of 114 features in total.

For the classification of a pair of mammograms according to a possible asymmetry, one-class classification is adopted, as described by Tax [36]. An example of the one-class classification scheme, using a 2-dimensional feature space, is shown in Fig. 10. One-class classification has been used in other applications successfully, e.g., in audio classification, described by Rabaoui et al. [37]. We train the one-class SVM classifier using the patterns of the asymmetric cases; then, we classify all the patterns using the trained classifier.

Using this classification scheme we try to model the class containing the asymmetric cases, as the patterns of this class tend to be close between themselves. All the symmetric cases can be considered as outliers and generally as non-asymmetric cases.

The features were processed through univariate significance analysis, specifically *T*-test, as stated by Cooley and Lohnes [38], resulting to a feature vector of pre-defined length of 18. For the one-class SVM classification, the libSVM software was used, as given by Chang and Lin [39]. For the kernel configuration, the radial basis function (RBF) was used. In order to test our system, the leave-one-out methodology was implemented, as described by Theodoridis and Koutroumbas [31].

4. Experiments and results

4.1. *Mammogram pre-processing*

The pre-processing techniques were applied to the images of the database. All the steps were successful, except for the image orientation algorithm, which failed in 3 of the images, where the breast was cut off, meaning that a large part of breast tissue is not included in the image, as Fig. 1b shows. However, this is a case of a non-acceptable mammographic

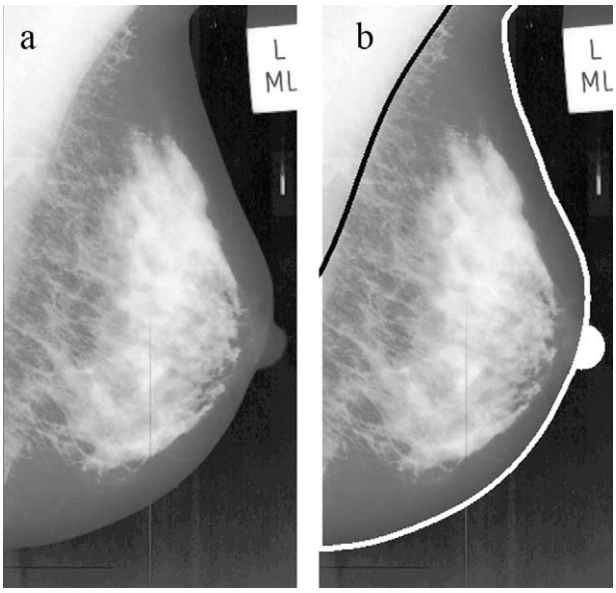

Fig. 11 – Segmentation scheme: (a) initial image and (b) breast boundary, pectoral muscle and nipple detected.

image. The noise is correctly detected and the tape artifacts are excluded from the subsequent processing of the image.

4.2. *Mammogram segmentation*

4.2.1. *Breast boundary detection*

The fully automatic breast boundary detection algorithm was tested on the images of the complete miniMIAS database. For the evaluation of the results, the images in Wirth [40] were used, as they correspond to the manual segmentation masks of the same images. The statistical measures adopted are the Tannimoto Coefficient (TC), as provided by Duda and Hart [41] and the Dice Similarity Coefficient (DSC), as proposed by Dice [42]. Considering two overlapping regions, A and B, these indices can be defined as $TC = (N(A \cap B))/(N(A \cup B))$ and $DSC = (2N(A \cap B))/(N(A) + N(B))$, having unity as the optimal value. A search area of 10mm around the "ground truth" boundary is defined, using the morphological operation of dilation and the TC and DSC metrics between the ground truth mask and the mask, obtained by the fully automatic breast border detection method, are estimated, at the search area defined before. In this way, only the region around the boundary is considered, so that to obtain a more reliable measure. We obtained the mean values of 0.900 and 0.945, for the TC and DSC respectively, for the 322 images of the database, whereas the corresponding standard deviations were 0.079 and 0.055. In other words, the fully automated segmentation algorithm gives significant results, similar to the manual segmentation method. An example is shown in Fig. 11b.

In order to ensure the fact, that the results of this stage are acceptable, we perform a direct comparison with the work of Wirth et al. [43]. There, a new algorithm for breast region segmentation using fuzzy reasoning was proposed. The evaluation is performed by comparing the extracted results with the same ground truth masks that we used. The metrics that are estimated in this work are *completeness*, *correctness* and *quality*, which are defined as $completeness = TP/(TP + FN)$, $correctness = TP/(TP + FP)$ and $quality = TP/(TP + FP + FN)$ correspondingly, where TP, FN and FP are the True-Positive, False-Negative and False-Positive pixels of the boundary. The mean values of the previous metrics of the results of the work of Wirth et al. [43] on the 322 images of the miniMIAS database were estimated as: $completeness = 0.996$, $correctness = 0.981$ and $quality = 0.980$. The corresponding values for the algorithm that we adopted are: $completeness = 0.993$, $correctness = 0.996$ and $quality = 0.989$. The obtained values are very similar and we will employ the method of Section 3.2.1 for the remaining processing steps.

4.2.2. *Pectoral muscle detection*

The pectoral muscle detection algorithm described in Section 3.2.2 was tested on the images of the database and the results were very promising. From Fig. 4, we can observe the output obtained via the already existing algorithm. Obviously, the initial straight line approximation (Fig. 4a) is refined to a more detailed estimate (Fig. 4b). However, the bottom end obtained is still not aligned with the actual boundary. Using our proposed modification, we achieve the estimate shown in Fig. 4c, which improves the detection of the boundary at this specific area. The difference of the pixels of the existing methodology and our approach is shown at Fig. 4d from which it is readily observer that our modified algorithm improves the estimate at the bottom end of the curve.

For the evaluation of the proposed algorithm, the following scheme is adopted: The existing algorithm is performed on the image i of the database, resulting to an estimate $P_{i,1}$. Then, our modified algorithm is performed on the same image, resulting to another estimate $P_{i,2}$. If the difference D_i of the two estimates is more than a specific number of pixels, *DiffPxls*, then the image is added to a set of images *DiffImgs*, which, at the end, corresponds to the images on which the proposed pectoral muscle detection algorithm gives significantly different results compared to the existing one. We chose to set the value of *DiffPxls* to 20 pixels and, as the image resolution used is 0.4mm/pixel, this results to an area of $E = 20 \times 0.4 \times 0.4 = 3.2\text{mm}^2$, which is an adequate and reasonable threshold area to differentiate between the two methods. At the end of this procedure, the *DiffImgs* set has a size of $N_D = 79$ images. All these images were given to an expert radiologist for evaluation. For each image i of this set, the radiologist gave a value to the variable $Mark_i$, according to the following marking scheme:

- $Mark_i = -2$, in case that the existing algorithm achieved surely a better pectoral estimate.
- $Mark_i = -1$, in case that the existing algorithm achieved a slightly better pectoral estimate.
- $Mark_i = 0$, in case that both algorithms succeeded or failed at the detection of the pectoral muscle.
- $Mark_i = +1$, in case that the proposed algorithm achieved a slightly better pectoral estimate.
- $Mark_i = +2$, in case that the proposed algorithm achieved surely a better pectoral estimate.

Then, the values of the metrics of average $a = (1/N_D)\sum_{i=1}^{N_D} Mark_i$ and weighted average $wa = (1/N_D)\sum_{i=1}^{N_D} D_i \cdot$

Table 3 – Truth table of the nipple detection algorithm.

Nipple	Not visible	Visible
Not detected	189	30
Detected	15	88

$Mark_i$ are estimated. The corresponding values are $a = 0.6329$ and $wa = 62.2$ pixels. The fact that both metrics are clearly positive, results to an evidence of the better performance of the proposed algorithm.

4.2.3. *Nipple detection*

For the evaluation of the nipple detection algorithm, expert radiologists annotated all the images in the database with regard to the visibility of the nipple. If it is in profile and visible, its exact location was given, using a customized user interface program. The proposed nipple detection algorithm, which should detect a nipple, in case it is in profile, was tested and the corresponding truth table is presented in Table 3. From the 118 mammograms, with a visible nipple, the nipple was correctly detected in the 88 of them, whereas in 30 mammograms no nipple was detected. These 30 mammograms were carefully observed and in 25 of them the nipple was recognized partly in profile (less than 1 mm). In these cases, the already detected breast boundary has succeeded in segmenting the nipple, i.e., including it inside the breast boundary itself. From the 204 mammograms with no nipple in profile, a nipple was detected to only 15 of them, resulting to false positive cases. After careful observation of these cases, it became apparent that the algorithm failed primarily due to the presence of very high noise levels.

An example of the results from the nipple detection process is shown in Fig. 11b. In order to evaluate the improvement obtained by the nipple detection technique, we estimate the new values of the TC and DSC measures, after including the detected nipple to the breast boundary estimation; their values were 0.903 and 0.947 with standard deviations 0.078 and 0.055, respectively. Although the increase with respect to values derived previously, is not large in absolute value, it must be noted that the boundary changes only in cases where the nipple is detected (103 images) and the area of the boundary change, due to the nipple presence, is too small compared to the whole breast boundary of the image.

4.3. *Mammogram classification*

4.3.1. *Breast density estimation*

The proposed mammographic breast density estimation algorithm was tested on all the images of the miniMIAS database, fully annotated according to the 3 breast density classes, as in Section 2 was explained. We preserved the initial classification scheme of 3 classes of the experts, in order to be able to compare directly with the algorithms of the literature. Note that masks capable of extracting the background, obtained by manual segmentation of the tissue-related areas given by Wirth [40], have been used. Thus, it was possible to compare the results derived by the fully automated and the manually segmentated techniques. For the evaluation of the algorithm, the work in Masek [15] was used, where the Closest Point Distance algorithm achieved 66.15% success rate, while a previous work of Blot and Zwiggelaar [44] reported 65%, when applied to a selected subset of the miniMIAS database. Both of these techniques use the leave-one-out methodology for the evaluation, so for the sake of a fair comparison we selected, in the first stage, this technique. For the classification step, we used several classifiers. The success rate for each one of the different classifiers, using the leave-one-out evaluation criterion is presented in Table 4. It is noteworthy that the SVM classification scheme outperformed by far the rest of the classifiers used, achieving a success rate of 85.71% for the automatic segmentation and 84.16% for the manual segmentation method. Table 5.

Table 4 – Classification results of each classifier for the breast density estimation step using the leave-one-out evaluation methodology.

Segmentation method	Classifier used		
	CART	k-nn	SVM
Automatic	67.39%	78.57%	85.71%
Manual	68.01%	78.57%	84.16%

Table 5 – Results of the proposed breast density estimation algorithm using the leave-one-out evaluation methodology. Values inside parentheses are the results obtained when using the manual segmentation method.

Breast density		True class		
		F	G	D
Predicted class	F	95 (92)	5 (9)	1 (1)
	G	11 (11)	89 (85)	19 (17)
	D	0 (3)	10 (10)	92 (94)

Table 6 – Classification results of each classifier for the breast density estimation step using the leave-one-woman-out evaluation methodology.

Segmentation method	Classifier used		
	CART	k-nn	SVM
Automatic	65.84%	76.40%	77.02%
Manual	65.84%	76.40%	77.33%

Table 7 – Results of the proposed breast density estimation algorithm using the leave-one-woman-out evaluation methodology. Values inside parentheses are the results obtained when using the manual segmentation method.

Breast density		True class		
		F	G	D
Predicted class	F	93 (94)	11 (15)	3 (4)
	G	12 (12)	72 (70)	26 (23)
	D	1 (0)	21 (19)	83 (85)

Table 8 – Results of the proposed asymmetry detection algorithm. Values inside parentheses are the results obtained when using the manual segmentation method.

Breast pair		True class	
		Symm.	Asymm.
Predicted class	Symm.	124(119)	3 (4)
	Asymm.	22(27)	12 (11)

Table 9 – Average processing time needed for each step of the algorithm for each mammogram of the database.

Step of the algorithm			Average processing time (s)		
1. Image preprocessing		1.1. Image orientation	0.3254		
		1.2. Noise estimation	1.7020		2.0394
		1.3.Image enhancement	0.0120		
2. Image segmentation		2.1. Breast boundary detection	5.0522		
		2.2. Pectoral muscle segmentation	8.2395		17.9203
		2.3. Nipple detection	4.6286		
3. Image classification	3.1. Breast density estimation	image preprocessing	0.0043		
		Feature extraction	1.0689	1.0885	
		Classification	0.0153		1.5583
	3.2. Asymmetry detection	feature extraction	0.4695	0.4698	
		Classification	0.0003		
	Processing time (all steps)			21.518	

The SVM classifier continues to achieve the best -among the other classifiers- success rates of 77.02% for the automatic and 77.33% for the manual segmentation methods, while the corresponding truth tables are provided at Table 7 provides the respective truth table of the SVM classifier; values inside the parentheses correspond to the manual segmentation method. When using the leave-one-woman-out evaluation technique, the corresponding results are presented in Table 6.

4.3.2. *Asymmetry detection*

The proposed asymmetry detection algorithm was applied to all the images of the miniMIAS database, which is fully annotated, by characterizing each pair of mammograms as symmetric (SYMM) or asymmetric (ASYMM). Breasts in mammograms are considered as symmetric organs; although they may differ in size, the internal structures are, usually, quite symmetric over broad areas of analysis. When true asymmetry (i.e., real 3-dimensional asymmetry, present in both projections, which is not the result of differences of positioning or compression) is present (either focal or global), it may be indicative of the presence of a mass or other abnormality, requiring further evaluation, as stated by Kopans [45]. The results of the algorithm are shown in Table 8. Similarly to the breast density estimation algorithm, the results derived by the fully automated and the manual segmentation techniques are presented together. The values, which are given in the parentheses are the corresponding results when using the manual segmentation technique. The evaluation of the algorithm is based on the work of Ferrari et al. [19], where an asymmetry detection technique using Gabor wavelets, described by Mallat [46], was presented and tested on a custom subset of 80 images of the miniMIAS database. The images were selected in such a way, so that to have equal number of symmetric and asymmetric cases and the average classification accuracy achieved was 74.4%. Moreover, the work of Rangayyan et al. [20] presents techniques to analyze bilateral asymmetry in mammograms by combining directional information, morphological measures, and geometric moments related to density distributions. The techniques were applied to 88 mammograms from the miniMIAS database, achieving classification accuracies of up to 84.4%. The results obtained using our new proposed algorithm were 80.75%, for the manual segmentation, and 84.47% for the automatic method. Note that our method is computationally much simpler and, more importantly, it is based on feature values that have already been computed and used in Section 4.3.1. Thus, our method addresses the tasks of mammographic breast density estimation and asymmetry detection in an automatic, unified and generic way.

4.4. *Processing time*

The processing time of all the previously reported steps is estimated and presented at Table 9. The experiments were carried out on a personal computer, equipped with an Intel Core 2 E6600 processor at 2.4GHz and 2GB of RAM. The software is developed on Matlab of version R2009b with no use of specific libraries or toolboxes. The average processing time of each mammographic image, for the whole described system, is 21.518 s; according to our experience, this time can be dramatically decreased if the system is written in a compiled programming language and using speed optimization techniques..

5. Discussion and conclusion

The complete system described in this paper and presented schematically in Fig. 12 was used for processing all the images of the miniMIAS database. All the intermediate results, i.e., breast boundary detection, pectoral muscle detection, nipple detection, asymmetry detection and breast density estimation, were examined in detail and evaluated by expert radiologists. It should be noted that the high level of noise, added to the images during the digitization process and the creation of the initial database images, makes the fully automated segmentation process a very challenging task.

The *pre-processing techniques*, which were selected to be applied in this work, were in general proved to be effective and successful, as the noise is correctly detected in most cases and sufficiently removed from the remaining stages of processing the images. The breast orientation algorithm failed

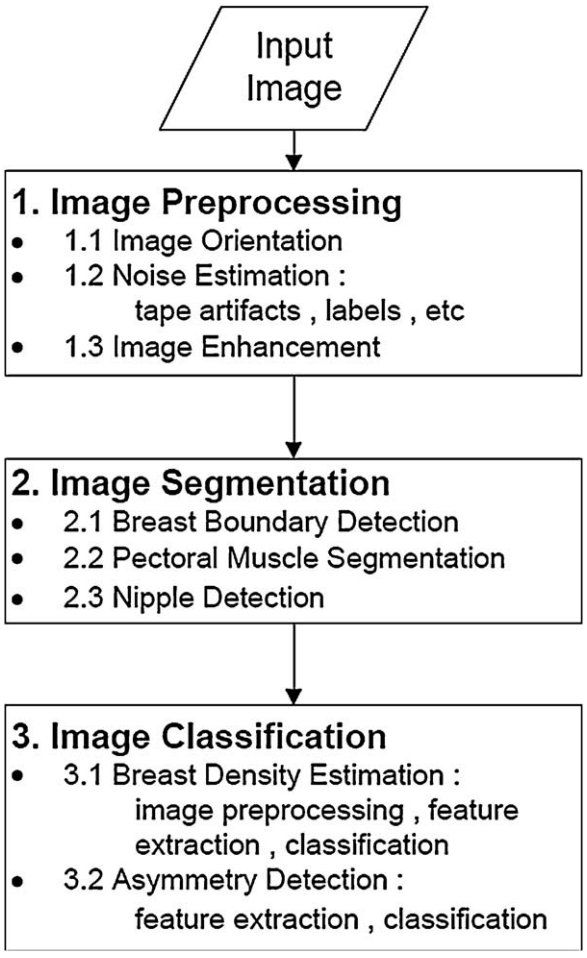

Fig. 12 – The complete system described.

in only three images, because in these cases the breast tissue is cut off from the image, as already explained in Section 4.1. These cases, however, are non-acceptable mammographic images, according to best practice and to the radiologist's opinion.

The implemented *breast boundary detection* technique, which is based on a simple inference, gives satisfactory results. This is obvious by a careful observation of the detected boundary of the images and also verified accordingly, using specific statistic measures. The *pectoral muscle estimate* is accurate and further improved, according to specific statistical measures, through the modification we propose. The new *nipple detection technique* tries to overcome the drawback of the breast boundary estimation method, i.e., not detecting the nipple, when this is in profile. In this way, it can serve as an improvement for the already established breast boundary, and in addition as a key point for further processing of the image, due to the importance of the nipple area in a mammographic image. Note that this technique can not be objectively compared to the algorithms proposed in the previously published relevant literature, since the most similar one is the work by Chandrasekhar and Attikiouzel [11], which uses only a small subset of the miniMIAS database and has a different target than ours. The results were evaluated by expert radiologists and are promising enough to expect even better results, when applied to high quality digital mammograms.

The proposed algorithm for *mammographic breast density estimation* achieves better results compared to the work of Masek [15] and of Oliver et al. [17], although the latter one uses only a selected small portion of the miniMIAS database. The work of Bosch et al. [16] achieves higher success rates, albeit it uses a different approach with higher-order textural features, which are computationally very expensive. The work we propose in this paper uses simple first-order statistical features and a new technique for the power spectrum estimation, making the whole process suitable for on-line training updates and real-time applications.

The *asymmetry detection* scheme uses the segmentation already obtained via the breast density estimation procedure. It achieves a success rate similar to or even higher than the levels reported in the relevant literature, although it uses the complete set of images of the miniMIAS database, instead of a small subset, as the work of Ferrari et al. [19] and Rangayyan et al. [20]. Therefore, our experimental results can be considered more reliable and consistent. Furthermore, the use of the one-class classification algorithm turned out to be a simple yet effective way to overcome the problem of the imbalanced classes, as the asymmetric cases are about 10% of the symmetric cases. The idea of the classification is to model as "target" the asymmetric cases and consider as "outliers" all the other cases, leading to an one-class scheme. The symmetric cases are not specifically modelled, but simply considered as non-asymmetric.

All the previously reported techniques can be combined and integrated to a clinical-level CAD system. All the algorithms are fully-automated and there is no need for external assistance. In addition, the processing time is not large enough, so each mammogram can be analyzed online; that is, on the fly as it is inserted the system. Moreover, the proposed scheme is considered to be robust against noise, as it has been verified by its application to the miniMIAS mammographic images database, in which the noise levels are very high and of varying nature.

Conflict of interest

There are no conflict of interest.

REFERENCES

[1] R. Nishikawa, Current status and future directions of computer-aided diagnosis in mammography, Computerized Medical Imaging and Graphics 31 (4–5) (2007) 224–235.
[2] M. Siddiqui, M. Anand, P. Mehrotra, R. Sarangi, N. Mathur, Biomonitoring of organochlorines in women with benign and malignant breast disease, Environmental Research 98 (2) (2005) 250–257.
[3] A. Wroblewska, P. Boninski, A. Przelaskowski, M. Kazubek, Segmentation and feature extraction for reliable classification of microcalcifications in digital mammograms, Opto-Electronics Review 11 (3) (2003) 227–235.
[4] R. Rangayyan, F. Ayres, J. Leo Desautels, A review of computer-aided diagnosis of breast cancer: toward the

detection of subtle signs, Journal of the Franklin Institute 344 (3–4) (2007) 312–348.
[5] R. Ferrari, R. Rangayyan, R. Borges, A. Frère, Segmentation of the fibro-glandular disc in mammogrms using Gaussian mixture modelling, Medical and Biological Engineering and Computing 42 (3) (2004) 378–387.
[6] V. Andolina, S. Lillé, K. Willison, Mammographic Imaging: A Practical Guide, Williams & Wilkins, Lippincott, 2001.
[7] S. Kwok, R. Chandrasekhar, Y. Attikiouzel, M. Rickard, Automatic pectoral muscle segmentation on mediolateral oblique view mammograms, IEEE Transactions on Medical Imaging 23 (9) (2004) 1129–1140.
[8] R. Ferrari, R. Rangayyan, J. Desautels, R. Borges, A. Frere, Automatic identification of the pectoral muscle in mammograms, IEEE Transactions on Medical Imaging 23 (2) (2004) 232–245.
[9] F. Yin, M. Giger, K. Doi, C. Vyborny, R. Schmidt, Computerized detection of masses in digital mammograms: automated alignment of breast images and its effect on bilateral-subtraction technique, Medical Physics 21 (1994) 445.
[10] H. Knauerhase, M. Strietzel, B. Gerber, T. Reimer, R. Fietkau, Tumor location, interval between surgery and radiotherapy, and boost technique influence local control after breast-conserving surgery and radiation: retrospective analysis of monoinstitutional long-term results, International Journal of Radiation Oncology, Biology, Physics 72 (4) (2008) 1048–1055.
[11] R. Chandrasekhar, Y. Attikiouzel, A simple method for automatically locating the nipple on mammograms, IEEE Transactions on Medical Imaging 16 (5) (1997) 483–494.
[12] A. Méndez, P. Tahoces, M. Lado, M. Souto, J. Correa, J. Vidal, Automatic detection of breast border and nipple in digital mammograms, Computer Methods and Programs in Biomedicine 49 (3) (1996) 253–262.
[13] J. Wolfe, Risk for breast cancer development determined by mammographic parenchymal pattern, Cancer 37 (5) (1976) 2486–2492.
[14] C. De Orsi, L. Bassett, S. Feig, et al., Illustrated Breast Imaging Reporting and Data System: Illustrated BI-RADS, American College of Radiology, Reston, VA, 1998.
[15] M. Masek, Hierarchical Segmentation of Mammograms Based on Pixel Intensity, Ph.D. Thesis, University of Western Australia School of Electrical, Electronic and Computer Engineering and University of Western Australia Centre for Intelligent Information Processing Systems, URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.128.1245, 2004.
[16] A. Bosch, X. Munoz, A. Oliver, J. Martı, Modeling and classifying breast tissue density in mammograms, in: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Volume 2, IEEE Computer Society Washington, DC, USA, 2006, pp. 1552–1558.
[17] A. Oliver, J. Freixenet, A. Bosch, D. Raba, R. Zwiggelaar, Automatic classification of breast tissue, in: Iberian Conference on Pattern Recognition and Image Analysis, Springer, 2005, pp. 431–438.
[18] M. Homer, Mammographic Interpretation: A Practical Approach, McGraw-Hill Companies, 1991.
[19] R. Ferrari, R. Rangayyan, J. Desautels, A. Frere, Analysis of asymmetry in mammograms via directional filtering withGabor wavelets, IEEE Transactions on Medical Imaging 20 (9) (2001) 953–964.
[20] R. Rangayyan, R. Ferrari, A. Frere, Analysis of bilateral asymmetry in mammograms using directional, morphological, and density features, Journal of Electronic Imaging 16 (1) (2007) 13003–113003.
[21] J. Suckling, J. Parker, D. Dance, S. Astley, I. Hutt, C. Boggis, I. Ricketts, E. Stamatakis, N. Cerneaz, S. Kok, et al., The mammographic image analysis society digital mammogram database, in: Exerpta Medica. International Congress Series, 1994, pp. 375–378.
[22] M. Mavroforakis, H. Georgiou, N. Dimitropoulos, D. Cavouras, S. Theodoridis, Significance analysis of qualitative mammographic features, using linear classifiers, neural networks and support vector machines, European Journal of Radiology 54 (1) (2005) 80–89.
[23] M. Masek, J. deSilva, Y. Christopher, Attikiouzel, Automatic breast orientation in mediolateral oblique view mammograms, in: 6th International Workshop on Digital Mammography, Greece, Springer-Verlag, 2003, pp. 207–209.
[24] A.D. Brink, N.E. Pendock, Minimum cross-entropy threshold selection, Pattern Recognition 29 (1) (1996) 179–188.
[25] I. Sobel, Camera Models and Machine Perception AIM-21, 1970.
[26] J. Radon, Über die Bestimmung von Funktionen durch ihre Integralwerte längs gewisser Mannigfaltigkeiten, Berichte Sächsische Akademie der Wissenschaften, Leipzig, Mathematisch-Physikalische Klasse 69 (1917) 262–277.
[27] R. Gonzalez, R. Woods, Digital Image Processing, Prentice Hall, 2007.
[28] A. Penn, M. Loew, Estimating fractal dimension with fractal interpolation function models, IEEE Transactions on Medical Imaging 16 (6) (1997) 930–937.
[29] L. Kaplan, Extended fractal analysis for texture classification andsegmentation, IEEE Transactions on Image Processing 8 (11) (1999) 1572–1585.
[30] L. Breiman, Classification and Regression Trees, CRC, Chapman & Hall, 1998.
[31] S. Theodoridis, K. Koutroumbas, Pattern Recognition, 4th edn., Academic Press, 2009.
[32] M.E. Mavroforakis, S. Theodoridis, A geometric approach to support vector machine (SVM) classification, IEEE Transactions on Neural Network 17 (3) (2006) 671–682.
[33] M. Mavroforakis, M. Sdralis, S. Theodoridis, A geometric nearest point algorithm for the efficient solution of the SVM classification task, IEEE Transactions on Neural Networks 18 (5) (2007) 1545–1549.
[34] H. Byun, S. Lee, A survey on pattern recognition applications of support vector machines, International Journal of Pattern Recognition and Artificial Intelligence 17 (3) (2003) 459–486.
[35] M. Mavroforakis, M. Sdralis, S. Theodoridis, A novel SVM geometric algorithm based on reduced convex hulls, in: Pattern Recognition, 2006. ICPR 2006. 18th International Conference, vol. 2, IEEE, 2006, pp. 564–568.
[36] D. Tax, One-class classification; Concept-learning in the absence of counter-examples, Ph.D. Thesis, 2001.
[37] A. Rabaoui, H. Kadri, Z. Lachiri, N. Ellouze, One-class SVMs challenges in audio detection and classification applications, EURASIP Journal on Advances in Signal Processing (2008) (19).
[38] W. Cooley, P. Lohnes, Multivariate Data Analysis, John Wiley & Sons Inc., 1971.
[39] C.-C. Chang, C.-J. Lin, LIBSVM: A Library for Support Vector Machines, Software available at http://www.csie.ntu.edu.tw/cjlin/libsvm, 2001.
[40] M. Wirth, MIAS Mask Database, University of Guelph, Canada, 2005.
[41] R.O. Duda, P.E. Hart, Pattem Classification and Scene Analysis, Wiley, New York, 1973.
[42] L. Dice, Measures of the amount of ecologic association between species, Ecology 26 (3) (1945) 297–302.

[43] M. Wirth, D. Nikitenko, J. Lyon, Segmentation of the breast region in mammograms using a rule-based fuzzy reasoning algorithm, ICGST International Journal on Graphics, Vision and Image Processing 05 (2005) 45–54.
[44] L. Blot, R. Zwiggelaar, Background texture extraction for the classification of mammographic parenchymal patterns, in: Medical Image Understanding and Analysis, 2001, pp. 145–148.
[45] D. Kopans, Breast Imaging, Williams & Wilkins, Lippincott, 2007.
[46] S. Mallat, A Wavelet Tour of Signal Processing, Academic Press, 1999.