A massive lesion detection algorithm in mammography


1. Dipartimento di Fisica e Tecnologie Relative dell’Università di Palermo e INFN - Catania (Italy)
2. Dipartimento di Fisica, Università di Napoli e INFN - Sezione di Napoli (Italy)
3. Struttura Dipartimentale di Matematica e Fisica, Università di Sassari e INFN - Sezione di Napoli (Italy)
4. INFN - Sezione di Torino (Italy)
5. Dipartimento di Informatica, Università di Torino e INFN - Sezione di Torino (Italy)
6. CEADEN, Havana, Cuba.
7. Dipartimento di Fisica dell’Università di Bari e Sezione INFN - Bari (Italy)
8. Dipartimento di Fisica dell’Università di Lecce e INFN - Sezione di Lecce (Italy)
9. Dipartimento di Scienza dei Materiali dell’Università di Lecce e INFN - Sezione di Lecce (Italy)
10. Center of Innovative Technologies for Signal Detection and Processing (TIRES), Bari (Italy)

Manuscript received: July 7, 2004; revised: October 29, 2004
Accepted for publication: November 26, 2004

Abstract

A new algorithm for massive lesion detection in mammography is presented. The algorithm consists in three main steps: 1) reduction of the dimension of the image to be processed through the identification of regions of interest (rois) as candidates for massive lesions; 2) characterization of the rois by means of suitable feature extraction; 3) pattern classification through supervised neural networks.

Suspect regions are detected by searching for local maxima of the pixel grey level intensity. A ring of increasing radius, centered on a maximum, is considered until the mean intensity in the ring decreases to a defined fraction of the maximum. The rois thus obtained are described by average, variance, skewness and kurtosis of the intensity distributions at different fractions of the radius. A neural network approach is adopted to classify suspect pathological and healthy pattern.

The software has been designed in the framework of the INFN (Istituto Nazionale Fisica Nucleare) research project GPCALMA (Grid Platform for CALMA) which recruits physicists and radiologists from different Italian Research Institutions and hospitals to develop software for breast and lung cancer detection.

keywords: Mammography, Neural Networks, CAD (Computer Aid Detection).

1. Introduction

Breast neoplasia is one of the most common kinds of tumour, besides the main cause for cancer mortality among women. A study made on 100,000 women between 1995-1998 in the European Union reports 39.4 deaths/year, regardless of the age, against 40.2 between 1985-1989, showing a change in rate equal to -2.1%. The favourable trend is due to therapeutic advancements, to screening and earlier diagnosis [1].

For this reason, screening programs are of great importance, as they enable early diagnosis in asymptomatic women which may greatly reduce the mortality.

Mammography [2, 3] is widely recognized as the most effective imaging modality for an early detection of breast cancer anomalies, and its proper scheduled execution is now the most used strategy for the earlier detection of this kind of tumour [4, 5].

The detection of anomalies in mammographic images [6] is made difficult by a great number of structures similar to the pathological ones, related also to tissue density. One of the abnormalities which is often a marker of a tumour is the presence of massive lesions, which are rather large objects with a diameter of the order of the centimetre and variable shapes.

In this paper, a new algorithm for an automatic search for massive lesions in mammographic images is described.

Mass detection plays a crucial role in the development of a Computer Aided Detection (CAD) system, and in recent years many studies have been made on this topic.

In 2004 Timp presented an automated technique to segment mass lesions from surrounding tissue with closed contours; with this method, based on dynamic programming, a value of $A_z$ (area under ROC curve) of 0.74 was obtained [7]. In 2003 Baydush proposed a study for detection of masses using subregion Hotelling observers: with this method a sensitivity of 98% and a specificity of 55.9% were obtained [8]. A template matching approach was developed by Tourassi obtaining a value of $A_z = 0.87 [9]$.

Address for correspondence: Francesco Fauci: e-mail: ffauci@unipa.it
CAD systems can be useful in screening programs as a huge amount of images has to be viewed by the radiologists. In a recent study Karssemeijer showed that double reading with CAD can improve sensitivity of masses up to 7.0% [10]; thus, accurate CAD routines are expected to provide valuable assistance to the radiologist in optimizing the sensitivity of the mammography examinations.

2. Materials and methods

2.1. Database

The database consists of 3,762 mammograms, belonging to 1,093 patients [11, 12], obtained by digitalization of the images collected in the hospitals of the GPCALMA collaboration. In particular 1,115 images, belonging to 608 patients, contain massive lesions. Statistical distributions on age and textures of the images present in the database are reported in Table I and Table II.

The films are digitized by means of a scanner with 85 micron pitch [13], and 12 bits, thus allowing 4096 grey levels. The uncompressed image size is about 12 Mbytes and in each examination, typically, four image are stored. All the images are characterized by a full description, e.g., radiologist description, tissue and lesion type, histological data. In particular, the pathological ones have a consistent description which includes radiological diagnosis and histological data, while non pathological ones correspond to patients with a follow up of at least three years.

Table I. Statistical distributions on age for database.

<table>
<thead>
<tr>
<th>&lt; 40 years</th>
<th>40 – 50 years</th>
<th>&gt; 50 years</th>
</tr>
</thead>
<tbody>
<tr>
<td>11 %</td>
<td>26 %</td>
<td>63 %</td>
</tr>
</tbody>
</table>

Table II. Statistical distributions on textures for database.

<table>
<thead>
<tr>
<th>Dense</th>
<th>Fibroadipose</th>
<th>Glandular</th>
</tr>
</thead>
<tbody>
<tr>
<td>6 %</td>
<td>36 %</td>
<td>58 %</td>
</tr>
</tbody>
</table>

It should be remarked that, as far as the image quality is concerned, the database is quite homogeneous: in fact, although the mammograms are recorded at different sites, their quality is almost the same for all images.

2.2. ROI Hunter

The identification of massive lesions requires a dedicated strategy: shape and dimension are often irregular and the borders are ill-defined, thus making difficult their discrimination from parenchymal structures [14, 15]. In Figure 1 some examples are reported.

Due to the large size of the mammographic image, a reduction of the surface under investigation, without loss of meaningful information, can be useful for the following steps. This can be achieved by means of a ROI (Region Of Interest) hunter algorithm, which should be able to select limited regions of the mammogram including peculiar characteristics as, for example, structures with a high density at the centre, sloping down with the distance.

To this purpose it is useful to define two parameters:

\[
T_N = \frac{I_{\text{pixel}} - I_{\text{average}}}{\sigma_{\text{average}}}; \quad T_R = \frac{I_{\text{ring}}}{I_{\text{max}}}
\]

where \(I_{\text{pixel}}\), \(I_{\text{average}}\), \(I_{\text{ring}}\), \(I_{\text{max}}\) and \(\sigma_{\text{average}}\) are defined in Table III.

Table III. Parameters definition.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>(I_{\text{pixel}})</td>
<td>intensity of a generic pixel</td>
</tr>
<tr>
<td>(I_{\text{average}})</td>
<td>average intensity of a mammogram</td>
</tr>
<tr>
<td>(I_{\text{ring}})</td>
<td>average intensity of a selected region (ring)</td>
</tr>
<tr>
<td>(I_{\text{max}})</td>
<td>intensity of a generic maximum</td>
</tr>
<tr>
<td>(\sigma_{\text{average}})</td>
<td>standard deviation of</td>
</tr>
</tbody>
</table>

An iterative procedure, based on the construction of a hierarchy of concentric rings centered on an intensity maximum (Fig. 2), has been implemented.

![Fig. 1. Some examples of massive lesions selected from the database.](image1)

![Fig. 2. Set of concentric rings, 5 pixel wide, drawn around the maximum intensity. The big dot is the center of the massive lesion.](image2)
to select the ROIs. The procedure consists in the following steps:

1. starting from the left top corner of the mammogram, a raster scanning of the image is performed to find an absolute intensity maximum (the centre of the candidate for the lesion);
2. if its intensity is lower than $T_N$ times the noise fluctuation around the average intensity of the mammogram, the algorithm stops, otherwise:
3. a set of concentric rings, 5 pixels wide, is drawn up to a maximum radius $R_{\text{max}}$ (that corresponds to the maximum expected size of a lesion, 250 pixels);
4. the pixel average intensity within each ring is computed;
5. the assigned ROI radius $R$ (with $R \leq R_{\text{max}}$) corresponds to the ring whose average intensity decreases to a $T_R$ fraction of the maximum;
6. the ROI is removed and stored for the following analysis; the corresponding “hole” left in the mammogram is set to zero;
7. a new maximum is searched in the remaining image; then go to step 2.

The values of the parameters $R_{\text{max}}$, $T_R$ and $T_N$ have been determined by running a dedicated software on the whole database. The results are shown in Figures 3a and 3b in terms of ‘efficiency’ (percentage of massive lesions correctly identified among those selected by the gold standard) and false positive per image (FP/im).

The goal of a ROI Hunter is to select all pathological regions but, unfortunately, a lot of non pathological regions are selected too. In order to discriminate these two kinds of regions, any CAD provides furthermore a classification tool, as the Neural Network; however, to get a better CAD performance, the ROI Hunter is requested to admit a low number of false positive. So, the best values of the above parameters have been set in order to maximize the efficiency and minimize the FP/image number. This means to choose the values at the knee in Figure 3a:

$$R_{\text{MAX}} = 2.4 \text{ cm} \quad T_R = 17\% \quad T_N = 6$$

by obtaining the overall efficiency:

$$\varepsilon = 86.2 \pm 1.2\%$$

and

$$\text{FP/image} = 9 \pm 1$$

As an example, in Figure 4 we report the original image (left), the extracted ROI (middle), and the image without the ROI (right).

2.3. Feature extraction

The ROI hunter provides «regions of interest» without giving further information. To this purpose suitable features should be selected such that a decision making system can correctly classify possibly
pathological regions from healthy ones. Different methods have been proposed to classify abnormalities in medical images using statistical parameters [16-19]. We consider the intensity distributions of the pixels belonging to a detected ROI of radius $R$. For illustrative purpose, the histograms of the grey level intensity are displayed in Figs. 5 and 6 for a typical region containing a massive lesion (Figures 7 - lesion A) and a typical region not containing a massive lesion (Figures 7 - lesion B) respectively; for both cases, we report the intensity histogram (scaled to 8 bit and normalized) inside the circle of radius $r = R$ (whole ROI, upper subfigure), $r = 2/3 R$ (middle) and $r = 1/3 R$ (down). It should be stressed that non pathological ROIs are likely to have flat intensity, giving rise to distributions which do not depend on the size of the region.

This can be clearly observed in Figure 6 where the spectrum does not shift or change appreciably at different values of $r$. Otherwise, in the presence of a pathological lesion, the selected region has an higher mean intensity value at the centre, which continuously decreases at increasing $r$; consequently, the spectrum shifts towards high luminosity values and it shape changes (see Figure 5).

The behaviour shown in Figures 5 and 6 can be properly analysed in terms of four statistical moments: average, variance, skewness and kurtosis (the last two features being indices of distribution asymmetry) of the intensity distribution as functions of $r$. For the case of a massive lesion, the average intensity increases, the variance decreases and the symmetry increases at decreasing $r$ (see Figure 5). This suggests the use of the values of the above mentioned statistical moments at $r = R$, $2/3 R$ and $1/3 R$ as suitable features for discriminating regions which contain massive lesions from regions which doesn’t contain massive lesions (see Figure 8). Similar behaviour is shown when the lesion is not perfectly centred by the ROI. Other decreasing steps can be used for $r$ (e.g., $r = R$, $3/4 R$, $1/2 R$, $1/4 R$), however no further performance improvement has been achieved. Although these features have been already used in other papers [20], we point out that in our approach we are able to mark the

---

**Fig. 4.** Original image (left), detected ROI (middle) and image without the ROI (right).

**Fig. 5.** Grey level intensity normalized distribution for a ROI with radius $R$, $2/3 R$, and $1/3 R$. The ROI refers to a pathological region.

**Fig. 6.** Grey level intensity normalized distribution for a ROI with radius $R$, $2/3 R$, and $1/3 R$. The ROI refers to a non pathological region.
possible pathological region by following the different shape of intensity distribution by decreasing the radius.

2. 4. Classification

The classification is performed by means of an Artificial Neural Network [21-23] with 12 input neurons (4 statistical moments previously described computed at three fractions of the roi radius, \( r = R, 2/3 \, R \) and \( 1/3 \, R \)), a number of hidden neurons which is tuned to obtain the best classification performance, and 1 output neuron.

The neural network gradually modifies the weights of the connections through a repeated presentation of a set of input patterns (training sample), with a prescribed answer, which is 1 for pathological patterns, 0 otherwise. This process is known as supervised learning. The weight modification is achieved through minimization of the difference between the neural output and the answer prescribed by the 'teacher'.

In our case, the original dataset is made of 980 pathological and 14,187 healthy patterns, extracted by the images related to 1,100 patients of the database, where pathological patterns correspond to the radiologist’s positive diagnosis, while healthy ones come from images without any lesion. The two samples are balanced by randomly extracting 980 patterns from the healthy sample. The dataset thus obtained was split in three subsets according to the following table:

<table>
<thead>
<tr>
<th>Pathological sample</th>
<th>Healthy sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training set</td>
<td>245</td>
</tr>
<tr>
<td>Validation set</td>
<td>245</td>
</tr>
<tr>
<td>Test set</td>
<td>490</td>
</tr>
</tbody>
</table>

The selected Neural Network is a feed-forward back-propagation supervised network trained with gradient descent learning rule with momentum, so as to quickly move along the direction of decreasing gradient, thus avoiding oscillations around secondary minima. The software is snns v4.1 (Stuttgart University) [23]. The weights were determined by training the network with the training set (Tab. IV) and minimizing the mean squared error (mse) of the validation set. The minimum mse value was achieved after 500 training epochs. Finally the performance of the overall system is evaluated on the test set.

More than one hidden layer may be useful for some applications, but one hidden layer is enough if a suitable number of hidden neurons is used [21]. The hidden layer size was optimised through a trial-and-error procedure and the minimum mse was obtained with 8 hidden neurons.
The neural output does not provide, in general, a binary value (true/false), rather a continuous value $0 \leq P \leq 1$ that may be used as threshold value to discriminate between regions with massive lesions and healthy regions. For each threshold value, sensitivity and specificity can be defined as follows:

$$\text{sensitivity} = \frac{N_{pp}}{N_p}$$
$$\text{specificity} = \frac{N_{hh}}{N_h}$$

where $N_{pp}$ and $N_{hh}$ are the number of correctly classified pathological and healthy patterns respectively, and $N_p$ and $N_h$ the total number of patterns belonging to pathological and healthy samples.

### 3. Results and discussion

The effectiveness of a diagnostic tool is provided by the ROC (Receiver Operating Characteristic) curve which is obtained by plotting, for different values of the threshold, the sensitivity against the False Positive Fraction (FPF), which is given by $(1 \cdot \text{specificity})$. The result obtained with this classification system is reported in Figure 9, where the ROC curve, obtained by neural analysis of the test set, is drawn. The area under the curve ($A_z$) may be used to evaluate the overall performance of the classification system. In our case, using the standard method described in [24-25], we obtain: $A_z = 76.9 \pm 1.3\%$.

In the following we summarize the main quantities describing the performance achieved by our algorithm:

- Efficiency = 86.2 ± 1.2%
- FP/image = 9 ± 1
- $A_z = 76.9 \pm 1.3\%$

Lesions which are missed by the ROI hunter are, mainly, spiculated masses, i.e., a particular kind of lesions resembling stars; this can be caused by the circular shape adopted to draw the ROI.

In order to get an overall CAD performance on the whole database the result are reported in term of FROC curve (Fig. 10) where the sensitivity is plotted vs FP/image.

It should be stressed that these results are insensitive to the particular kind of tissue texture (dense, fibroadipose and glandular). The computing time is of the order of 5 seconds for each image.

### 4. Conclusion

In this paper a new algorithm for massive lesion detection has been presented. The algorithm is based on a ROI hunter procedure able to select regions with higher probability to contain massive lesions. The selected ROIs are described by four statistical moments, i.e., average, variance, skewness and kurtosis of the grey level intensity distribution, at three different fractions of the ROI radius. Through these features have already been used in similar approaches, it should be remarked that we exploited their trend over the ROI portion as a clear mark of the pathology. The discriminating performance of the algorithm was checked by means of a supervised neural network, and the results in terms of ROC curve have been presented.

Results in terms of sensitivity, specificity and $A_z$ are comparable to that obtained in other recent studies [8-10], verifying the integrity of our approach.

The choice of a circle to delimit the ROI is a rough approximation of the shape of a massive lesion which implies undesired noise in the features extraction when the lesion has not a spherical symmetry. Other segmentation methods are under study to overcome this limitation.

The algorithm has been designed in the framework of the INFN GPCALMA research project which aims at developing CAD systems for early diagnosis.
Fig. 11. Layout of the gpcalma station displaying both CAD (dashed line) and radiologist’s (continuous line) diagnosis.

of breast cancer. Figure 11 shows a typical layout of the software displaying both CAD (dashed line) and radiologist’s diagnosis (continuous line). It is to stress that any CAD system should be intended to support radiologist’s diagnosis not to replace it.

Acknowledgement

We acknowledge the help of the staff and the institutions involved in the study. In particular, we thank Prof. V. Lattanzio (Policlinico di Bari), Prof. Bazzocchi (Dipartimento di Ricerche Mediche e Morfologiche della Università, Udine, Italy), Prof. E. Zanon (Ospedale Valdese di Torino and INFN, sezione di Torino), Prof. A. Sodano (Dipartimento di Scienze Biomorfologiche e Funzionali Università di Napoli “Federico II” Napoli, Italia) and Prof. R. Lagalla, Dr. R. Lenzi, Dr. G. Bellissima, Dr. V. Ricotta (Istituto di Radiologia P. Cignolini – University of Palermo) for the medical support in the data acquisition.

References


F. Fauci et alii: A massive lesion detection algorithm in mammography


