Ciência eNatura, v. 37 Part 2 jun. 2015, p. 139–145

ISSN impressa: 0100-8307 ISSN on-line: 2179-460X

ciência e natura

Fusion of Hidden Markov Model and Wavelet Transform to Recognise Breast Cancer

Sayedeh Somayeh Hosaini[1,*], Mehran Emadi[2]

[1]Computer Master Student, Department of Computer Engineering,, Najafabad Branch, Islamic Azad University, Najafabad, Iran; Yass957@yahoo.com
[2]Assistant professor, Department of Electrical Engineering,, Mobarakeh Branch, Islamic Azad University, Mobarakeh, Iran; m.emadi@mau.ac.ir

Abstract

Public health is affected by the health and vitality of women who are the pillars of the families. Among the women between 15 to 54 years old, breast cancer is one of the most important causes of mortality. Every 13 minutes, a woman loses her life due to breast cancer. Statistics show that %12.6 of women suffer from breast cancer. Mammography can be considered as one of the most efficient methods of breast cancer diagnosis; however, this method has still its own limitations. The quality of mammography images can be enhanced by using image processing and the disease can be diagnosed more easily by using image feature extraction. Using Hidden Markov Model and wavelet transform method, a new method for the detection of suspicious areas of breast cancer tumors will be proposed in this study. In addition to the detection of tumors, this method can determine the mass percentage to estimate the progression of the disease. In this study, Markov Model with a tree structure was used to extract statistical properties of the wavelet components. The specific capacity of Markov Model to extract information about edges and the regions of protrusions in image tissues increases the accuracy of cancerous areas detection. The present study sought to estimate the appropriate label (clustering) of the pixels of an image under investigation to segment cancerous areas. Therefore, certain joint distributions were assumed for the pixels of a region or a class and then the maximum similarities between different areas of the examined image and joint distributions were examined through the Maximum-Likelihood Estimation method. To evaluate the proposed method and analyze the results, a combination of the MIAS and PADN databases comprising 150 images was used. The results indicated that the proposed method is more accurate than methods which are solely based on wavelet transform. In the new method, the obtained detection rate was %96 while in the wavelet transform method, it has been reported %71.5 indicating the precision of the proposed method.

Keywords*: image processing, mammograms, breast cancer, wavelet transform, Hidden Markov Model, distribution function.*

Recebido: dia/mês/ano Aceito: dia/mês/ano

1 Introduction

Among the women between 15 to 54 years old, breast cancer is one of the most important causes of mortality. Statistics show that %12.6 of women suffer from breast cancer. In other words, every 13 minutes, a woman loses her life due to breast cancer [1].

Early diagnosis is an important factor in the treatment and recovery of breast cancer that increases the chances of survival. The most common method of breast cancer diagnosis is surgical biopsy. Among all available methods of breast cancer diagnosis, surgical biopsy is the most accurate and time-consuming method [2]. Mammography is another method of breast cancer diagnosis. Through this method, masses and symptoms of cancer can be detected ten years before they become tangible [3]. A mammogram is an x-ray image of the breast tissue by which a radiologist can detect abnormal symptoms such as lumps, calcification, structural breakdown and breast asymmetry [4].

Although mammography is a perfect tool for early diagnosis of breast cancer, it is still possible that an injury be interpreted as cancer or cancer be misdiagnosed due to available limitations in the process of mammogram interpretation. As a result, the rate of misinterpretation by the radiologists is 10-30 percent in this method [5]. When an injury is wrongfully diagnosed as cancer, unnecessary biopsy and surgery will be carried out on the patient while wrongfully ignored cancer may lead to the patient's death.

Image processing techniques can be used to improve the diagnosis and enhance its sensitivity and accuracy. Image processing is a technique for converting an image to digital form and performing some operations on it in order to obtain an improved image or to extract some useful information from it.

In the present study, image processing technique was used to intensify and improve mammogram images and a combination of wavelet transform method and Hidden Markov Model (HMM) was applied to increase the accuracy and efficiency of breast cancer tumors detection. Furthermore, 150 mammogram images were collected from the MIAS and PADN databases and radiology offices for the examination of the proposed method of breast cancer detection.

The MIAS database has been collected by the Mammographic Image Analysis Society in the United Kingdom. In this database, the size of all images is 1024*1024. The MIAS contains 322 mammograms of 161 women's left and right breasts with a resolution of 50 micron pixel edge (each pixel is represented with an 8-bit gray depth). Pathologically, this database include 209 images of normal breast, 67 images of breasts with benign (non-cancerous) tumors and 54 images of breasts with malignant (abnormal) tumors [6].

The PADN database has been collected in Singapore and includes 150 mammogram images from a variety of benign and malignant breast tumors. In this database, images have different resolutions; but, their size is 600*800 pixels.

Life expectancy in patients and treatment of breast cancer are subject to timely diagnosis of the disease; thus, early diagnosis of breast cancer is a priority. Accordingly, doctors and radiologists have applied the science of image processing to diagnose breast cancer tumors early. In the following section, the related literature will be reviewed. In 2008, Papadopoulos and colleagues examined the methods of Contras Limited Adaptive Histogram Equalization (CLAHE), Local Maximum Region (LMR), Redundant Discrete Wavelet (RDW) and linear elasticity algorithm to improve tumor diagnosis through mammograms. They used images obtained from the MIAS and Nijmegen Digital Mammogram databases and found the LRM as the best method for breast cancer tumor diagnosis [7]. Using thresholding method, Bhadoria and colleagues first removed the parts of normal breast tissue from the image and then analyzed the growth of the breast tumor and its characteristics using the shape of the tumor [8]. In 2011, Maitra and colleagues offered a method to detect abnormal masses in mammogram images in which the image is divided into two separate parts with the formation of homogeneous blocks and

Recebido: dia/mês/ano Aceito: dia/mês/ano

determination of color after processing. To evaluate their offered method, they used the MIAS database and reported satisfactory results including obtaining the results as soon as possible and the simplicity of the method [9]. To increase the resolution of tumors in mammogram images, Natarajan and colleagues removed noises from the images and then detected tumors more precisely through contrast adjustment, image stretching, subtraction, threshold development and segmentation [10]. In 2014, Ojo and colleagues proposed an algorithm to effectively remove noises from mammogram images by using splitting and merging techniques. To evaluated their method, they used 322 images obtained from the MIAS database and could adequately diagnose the problems in 297 images (%92.24).This method reduces false positives in the diagnosis of breast cancer. [11]. In 2015, Kanchana and colleagues detected breast cancer tumors in mammogram images by using a wavelet-based method and thresholding. In their proposed method, wavelet transform was used based on four levels of analysis and mammogram images were analyzed using statistical properties of wavelet and vector target. The evaluation of their method on the MIAS database represented %92.30 accuracy in breast cancer diagnosis [12].

The overall process of breast cancer detection used in this study is presented in Figure (1).

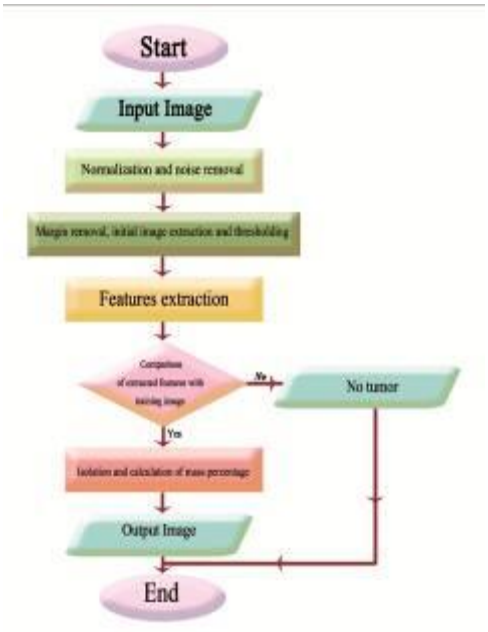


Fig.1. The overall process of breast cancer diagnosis used in this study

2 Methodology

2.1 Preprocessing

At the first step, the preprocessing operations of noise removal, smoothing and margin removal were done on the mammogram image. For smoothing, the median filter was used. For edging, the Sobel Operator was used to calculate the gradient of pixels in a neighborhood. For margins removal and accurate detection of the main areas, edges were found first and then, the edge image with 0 and 1 values was profiled both vertically and horizontally. Therefore, with the selection of a small threshold, top and bottom margins were identified and removed. The same procedure was conducted for left and right margins as well. An example margin removal is presented in figure (2).

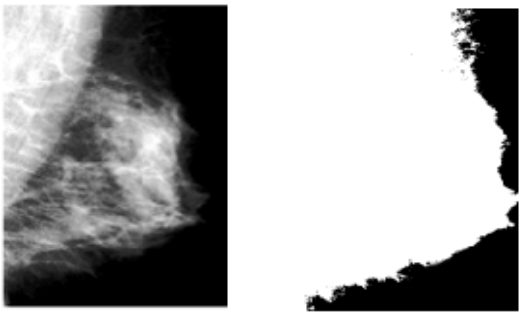


Fig.2. The initial mammogram image and the removed margin from the initial image

3.2 Processing

For image segmentation and detection of cancerous regions, the process can be done a step at a time and based on a specific window size; so that the dependency of pixels in each window on a class can be checked. Notably, the determination of window size is very important. If windows with greater size are used for the segmentation, the accuracy increases because more pixels will be examined and better statistical information will be provided. However, the risk of having pixels of other classes in the same window will increase. Thus, greater windows provide more precise segmentation in large and homogeneous areas;

Recebido: dia/mês/ano Aceito: dia/mês/ano

but, less accurate segmentation will be the result in the border areas. On the other hand, small windows decrease the chance of having pixels of different classes in the same window; but, they cannot perform an accurate segmentation due to a limited number of statistical data. Therefore, it is better to use small windows in the border areas and great windows in other parts of an image. In this study, the dyadic squares (or blocks) were employed to implement segmentation windows. To do so, the selected windows for segmentation were divided into four square sub-sections recursively at each stage. With this procedure, the images under consideration would be segmented based on the class label estimation for each of the windows. That estimation also required a Probability Distribution Function (PDF) for pixels of each class. The important issues in this process were wavelet-based transformation and statistical method.

It is very important to have models of different tissues in the segmentation of suspicious areas because access to a perfect joint distribution of pixels is very difficult. Thus, the basis of different regions analysis in this study was to convert domains using wavelet because that linear convertible transformation led to the emergence of components which were less complicated to be modeled. Most of the images, especially the gray ones, have a structure (bumps and edges) through which wavelet transform leads to an accurate segmentation. In fact, wavelet transform in these cases operates as multi-scale edge detection explaining the structure of an image based on different scales and in three different directions. For edges, wavelet transform leads to the emergence of major components and for other areas, it leads to the emergence of minor components.

Many statistical methods have been proposed to model tissue structure. Among all these models, the HMM was applied in the present study. The HMM was used to approximate the marginal and joint statistics of wavelet components. Through the HMM, a variable of hidden status is considered for each of the wavelet components to control their largeness or smallness. Then, the marginal distribution of each component will be modeled through the combination of two Gaussian distribution. It must be noted that in the modeling of large areas, large variances and in the modeling of small areas, small variances are used for Gaussian distributions. It is also noteworthy that two Gaussian distribution is capable of modeling marginal statistical distribution of non-Gaussian wavelet components in the actual images. In the course of scale changing, the HMM extracts the persistency of large and small components by using the dependency between hidden statuses.

For any mammogram image under investigation, a tree window structure was created first and then each window's wavelet transform for data extraction was calculated. After that, each wavelet component was modeled based on the two Gaussian distribution and its variance was controlled by the variable of hidden Markov status.

At this stage, parameters of the model, which was a combination of the two Gaussian distribution variances and the Markov transition probabilities, were adapted into an M vector. Thus, the HMM can be used as a model for the approximation of all joint distributions of the w wavelet components f(w/M). To estimate the parameters of the HMM, the iterative method of EM can be used as an example. The three HMM has an interesting structure to which all windows can be adapted. Each sub-tree of a tree HMM is another tree HMM which starts at the ith node and models the statistical behavior of wavelet components related to the dith window. Therefore, the distribution of each window in the process of image modeling is as follows:

$$f(d_i|M) \quad (1)$$

Then, the process of image segmentation could be started. It was assumed that for each tissue class $c \in \{1,2,\dots,N_c\}$, a three HMM was trained based on the MC parameters. Then the w wavelet transform of a training image x including those tissues was considered. Therefore, the calculation of multi-scale distribution expressed for different windows led to the emergence of $f(\tilde{d}_i|M_c), c \in \{1,2,\dots,N_c\}$ for each of the dyadic sub-images di. At this stage, segmentation of the image under consideration could be done in terms of maximum similarity criterion as follows:

$$\hat{c}_i^{ML} = \arg_{c\in\{1,2,\dots,N_c\}} \max\ f(\tilde{d}_i|M_c) \quad (2)$$

This segmentation process created a set consisting of J members of various segmentations:

Recebido: dia/mês/ano Aceito: dia/mês/ano

$$\hat{c}_i^{ML}, j = 0,1,\dots,J-1$$

4. Evaluation

To evaluate the proposed method, mammogram images obtained from the MIAS and PADN databases were used. First, an image of a cancerous tissue was selected and used for the tree HMM training. Then, an image (figure.3) was considered for the segmentation using the above mentioned mechanism and a tree structure of the windows:

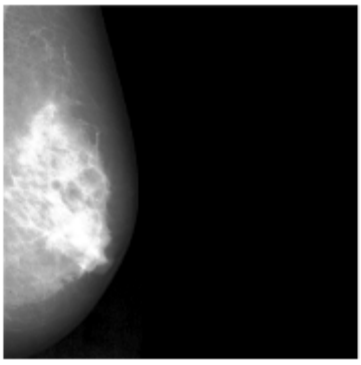


Fig.3. The training image

The result of tumor separation based on the proposed method is presented in figure (4).

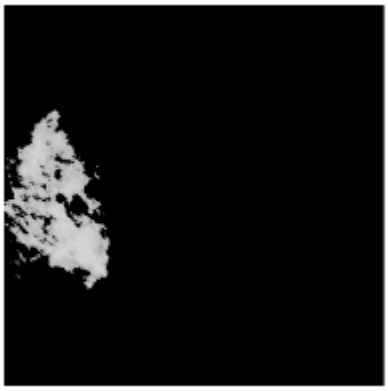
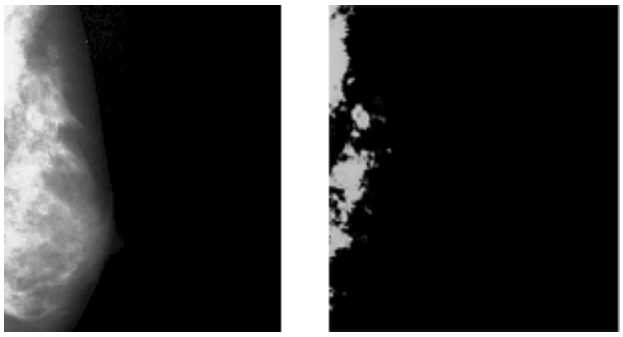


Fig.4. Result of tumor separation based on the proposed method

To estimate the growth of the tumor, periodical mammography was done and the mass percentage was calculated and reported. For example, the calculated mass percentage in the following image is %12.4127.



A B

Fig.5. Mammogram image before processing (A) and the extracted tumor after processing (B)

In the following sections, the performance of the proposed method is compared with the interpretation of mammography results by an expert. The results obtained from the proposed method are presented in table (1).

Table 1. Results obtained from the proposed method

Sample/number of samples	Result
Simple Sample/98	100 %
Complex Sample/52	92%
Total/150	96%

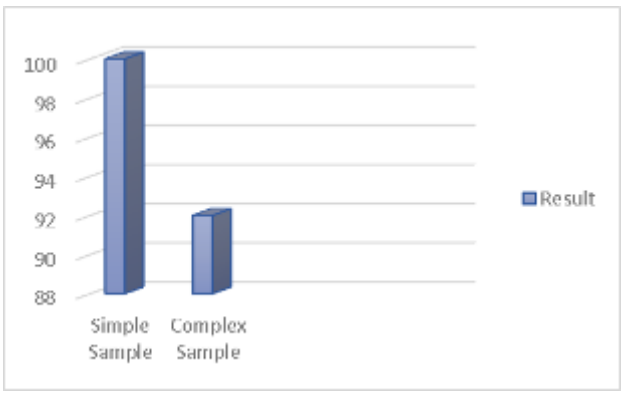


Fig.6. Results obtained from the proposed method (%)

As the results showed, the proposed method was able to detect all cases of tumors in simple cases of mammography and regarding the more complicated cases, the proposed method could detect %92 of the cases. The overall detection rate of %96 was obtained using the proposed method. The proposed method utilized wavelet transform and the HMM to detect breast cancer

Recebido: dia/mês/ano Aceito: dia/mês/ano

tumors in mammogram images and separate them from the images. The results also showed that the newly proposed method provided more accurate detection of breast cancer tumors compared to the method in which only wavelet transform is used.

The differences between the two methods are presented in table (2).

Table 2. Comparison of the results obtained from the proposed method and the wavelet method

Sample/number of samples	Result of Wavelet & Hidden Markov Model	Result of Wavelet
Simple Sample/98	100 %	82 %
Complex Sample/52	92%	61%
Total/150	96%	71.5%

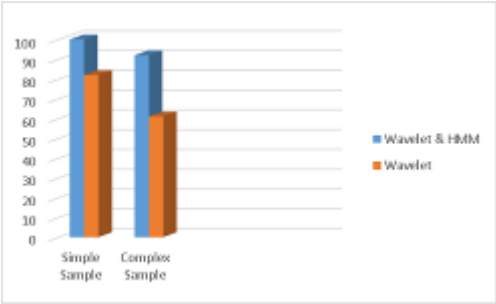


Fig.7. Comparison of the results obtained from the proposed method and the wavelet method (%)

5 Conclusion

Despite great advances in the field of mammography in the past two decades, a great number of women die of breast cancer. Mammography has always been one of the most reliable ways to diagnose this disease. This study aimed to propose a new reliable method of breast cancer diagnosis using a combination of wavelet transform method and the HMM. Therefore, Markov Model with a tree structure was used to extract statistical properties of the wavelet components. The specific capacity of Markov Model to extract information about edges and the regions of protrusions in image tissues increases the accuracy of cancerous areas detection. To do this, a training image containing a definitively diagnosed breast cancer tumor was used. According to the new method, all features of the training image were extracted and recorded. Similarly, the features of the input mammogram image was obtained and compared with the features of the training image through the Maximum-Likelihood Estimation method. Finally, the existence or non-existence of tumor in the input image was announced. In addition to the detection of tumors, this method could determine the mass percentage to estimate the progression of the disease allowing the specialists and radiologist to diagnose the disease more confidently. To evaluate the proposed method and analyze the results, a combination of the MIAS and PADN databases comprising 150 images was used. The results indicated that the proposed method was more accurate than methods solely based on wavelet transform. In the new method, the obtained detection rate was %96 while in the wavelet transform method, it has been reported %71.5 indicating higher precision of the proposed method.

References

World-Health-Organization. (2013). Cancer, WHO, Fact sheet N°297. Available: http://www.who.int/mediacentre/factsheets/fs297/en/

M. Fiyouzi, K. Rezaei, J. Haddadnia, “A new method of breast cancer diagnosis and extraction of breast cancer tumors from mammogram images,” Fourteenth Student Conference on Electrical Engineering Iran, University of Kermanshah, September, 2011.

http://www.dralighayour.com.

B. Verma and P. Zhang, "A novel neural-genetic algorithm to find the most significant combination of features in digital mammograms," Applied Soft Computing, vol. 7, pp. 612-625, 2007.

M. L. Giger, "Computer-aided diagnosis in radiology," Academic Radiology, vol. 9, pp. 1-3, 2002.

http://peipa.essex.ac.uk/info/mias.html.

Recebido: dia/mês/ano Aceito: dia/mês/ano

A. Papadopoulos, D. I. Fotiadis and L. Costaridou, "Improvement of microcalcification cluster detection in mammography utilizing image enhancement techniques", Computers in biology and medicine, 1045-1055, 2008.

S. Bhadoria, C. Dethe, S. Patra and A. Chaubey, "Breast Tumor Shape and Size Analysis for Growth Estimation Using Mammography", International Journal on Advanced Computer Engineering and Communication Technology.

I. K. Maitra, S. Nag and S. K. Bandyopadhyay, "Identification of abnormal masses in digital mammography images", International Journal of Computer Graphics, 17-30, 2011.

P. Natarajan, D. Ghosh, K. N. Sandeep and S. Jilani, "Detection of Tumor in Mammogram Images using Extended Local Minima Threshold",International Journal of Engineering & Technology (0975-4024), 2013.

J. Ojo, T. Adepoju, E. Omdiora, O. Olabiyisi and O. Bello, "Pre-Processing Method for Extraction of Pectoral Muscle and Removal of Artefacts in Mammogram" 2014.

M. Kanchana, P. Varalakshmi, "Breast Cancer Diagnosis Using Wavelet Based Threshold Method",ISSN 1990-9233,2015.

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

Recebido: dia/mês/ano Aceito: dia/mês/ano
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~