International Journal of Bioinformatics Research
ISSN: 0975–3087 & E-ISSN: 0975–9115 , Volume 4, Issue 1, 2012, pp.-241-244
Available online at http://www.bioinfo.in/contents.php?id=21

FEATURE EXTRACTION OF MAMMOGRAMS

PRADEEP N.[1*], GIRISHA H.[2], SREEPATHI B.[2] AND KARIBASAPPA K.[3]

[1]Department of Computer Science and Engineering, B.I.E.T., Davangere, Karnataka, India.
[2]Department of Computer Science and Engineering, RYMEC, Bellary, Karnataka, India.
[3]Department of Computer Science and Engineering, DSCE, Bangalore, Karnataka, India.
*Corresponding Author: Email- nmnpradeep@yahoo.com

Received: January 20, 2012; Accepted: February 20, 2012

Abstract- Cancer is uncontrolled growth of cells. Breast Cancer is the uncontrolled growth of cells in the breast region. Breast cancer is the second leading cause of cancer deaths in women today. Early detection of the cancer can reduce mortality rate. Early detection of Breast Cancer can be achieved using Digital Mammography, typically through detection of characteristic masses and/or microcalcifications. A Mammogram is an x-ray of the breast tissue which is designed to identify abnormalities. Studies have shown that radiologists can miss the detection of a significant proportion of abnormalities in addition to having high rates of false positives. Therefore, it would be valuable to develop a computer aided method for mass/tumor classification based on extracted features from the Region Of Interest (ROI) in mammograms. ROI has to be segmented from the digital mammogram using the Segmentation techniques. Pattern recognition in image processing requires the extraction of features from regions of the image, and the processing of these features with a pattern recognition algorithm. We consider the feature extraction part of this processing, with a focus on the problem of tumor detection in digital mammography.
Features are nothing but observable patterns in the image which gives some information about the image. For every Pattern Classification problem, the most important stage is Feature Extraction. The accuracy of the classification depends on the Feature Extraction stage. The different features that can be extracted for a digital mammogram are: Texture Features, Statistical Features, Structural Features.
In this paper, we are able to calculate Texture, Statistical and Structural Features. We have used MATLAB for extracting the tumors from input mammogram and for calculating various features.
Keywords- Breast Cancer, Digital Mammography, Region of Interest, Segmentation, Features, Texture, Statistical, Structural

Citation: Pradeep N., et al (2012) Feature Extraction of Mammograms. International Journal of Bioinformatics Research, ISSN: 0975–3087 & E-ISSN: 0975–9115 , Volume 4, Issue 1, 2012, pp-241-244.

Introduction

Breast cancer remains a leading cause of cancer deaths among women in many parts of the world. Early detection of breast cancer through periodic screening has noticeably improved the outcome of the disease [1]. Cancer is an abnormal, continual multiplying of cells. The cells divide uncontrollably and may grow into adjacent tissue or spread to distant parts of the body. The mass of cancer cells will eventually become large enough to produce lumps, masses, or tumors that can be detected. Breast Tumor is a tumor present in Breast. Tumor is uncontrolled growth of cells which can be either benign or malignant. Benign is not cancerous. Benign tumors may grow larger but do not spread to other parts of the body. Malignant is cancerous. Malignant tumors can invade and destroy nearby tissue and spread to other parts of the body. Tumor can be easily identified in mammogram because tumor part is highly bright (having high intensity) compared to other part (background) of the mammogram image as shown in "Fig. (1)".

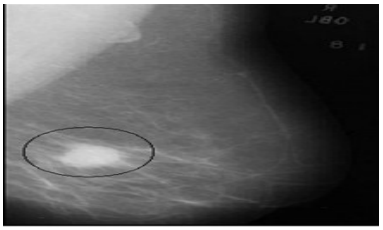

Fig. 1- Sample mammogram

In the figure, we can observe that the marked oval shape area have higher intensity compared to the surrounding area. This marked oval shape is the required ROI. The most important and challenging task is to segment only the tumor from the digital mammogram. Features have to be determined for the segmented tumor. Computer Aided (CA) detection systems have been developed to aid radiologists in detecting mammographic lesions, charactering promising performance [2-5].

There are large numbers of diagnostic methods currently available, among which mammography is the most reliable method, for detecting early breast cancer [6-7].

Methodology

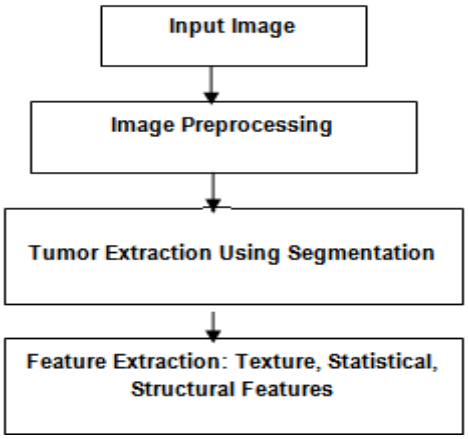

Fig. 2- Methodology

Image Preprocessing

The preprocessing phase of digital mammograms refers to the enhancement of mammograms intensity and contrast manipulation, noise reduction, background removal, edges sharpening, filtering, etc.

Segmentation

In analyzing mammogram image, it is important to distinguish the suspicious region from its surroundings. The methods used to separate the Region of Interest from the background are usually referred as the segmentation process. Segmentation can be carried out using any of the standard techniques like Local Thresholding, K –Means Clustering, Otsu Segmentation Technique. In this paper, we have used Local Thresholding Technique [8] for segmentation. The segmentation block diagram is shown in the "Fig. (3)".

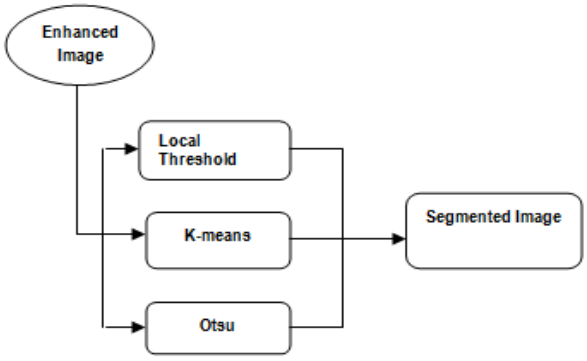

Fig. 3- Segmentation block diagram

Local Thresholding

The technique has been proven to provide an easy and convenient way to perform the segmentation on digital mammogram. The segmentation is determined by a single value known as the intensity threshold value. Then, each pixel in the image is compared with the threshold value. Pixel intensity values higher than the threshold will result in a white spot in the output image

Feature Extraction

Feature is used to denote a piece of information which is relevant for solving the computational task related to a certain application. More specifically, features can refer to:

- The result of a general neighborhood operation (feature extractor or feature detector) applied to the image,
- Specific structures in the image itself, ranging from simple structures such as points or edges to more complex structures such as objects.

Many features have been extracted for the abnormalities of mammograms. The extraction methods of texture feature play very important role in detecting abnormalities of mammograms because of the nature of mammograms. Texture features have been proven to be useful in differentiating masses and normal breast tissues. Texture features are able to isolate normal and abnormal lesion with masses and microcalcification. Feature extraction block diagram is shown in "Fig. (4)".

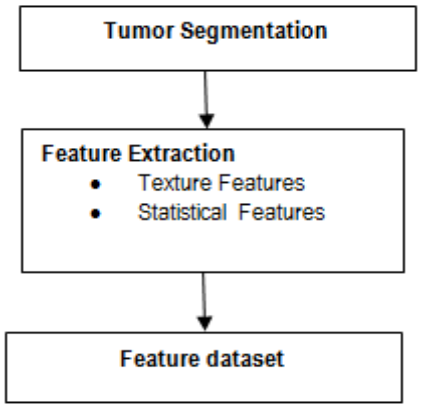

Fig.4- Feature Extraction Block diagram 1

In the work, we have extracted Texture features, Statistical features and Structural Features for the segmented tumor from the given input mammogram image. The features that we have extracted are:

1. Mean

The mean, m of the pixel values in the defined window, estimates the value in the image in which central clustering occurs. The mean can be calculated using the formula:

$$\mu = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N} p(i,j)$$

Where p(i,j) is the pixel value at point (i,j) of an image of size MxN.

2. Standard Deviation

The Standard Deviation, σ is the estimate of the mean square deviation of grey pixel value p (i, j) from its mean valuem. Standard

deviation describes the dispersion within a local region. It is determined using the formula:

$$\sigma = \sqrt{\frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}(p(i,j)-\mu)^2}$$

3. Smoothness

Relative smoothness, *R* is a measure of grey level contrast that can be used to establish descriptors of relative smoothness. The smoothness is determined using the formula:

$$R = 1 - \frac{1}{1+\sigma^2}$$

Where, σ is the Standard Deviation of the image.

4. Entropy

Entropy is a statistical measure of randomness that can be used to characterize the texture of the input image. Entropy, h can also be used to describe the distribution variation in a region. Overall Entropy of the image can be calculated as:

$$h = -\sum_{k=0}^{L-1} Pr_k(log_2 Pr_k)$$

Where, Pr is the probability of the k-th grey level, which can be calculated as Zk /m*n, Zk is the total number of pixels with the kth grey level and L is the total number of grey levels.

5. Skewness

Skewness, S characterizes the degree of asymmetry of a pixel distribution in the specified window around its mean. Skewness is a pure number that characterizes only the shape of the distribution. The formula for finding Skewness is given in the below equation:

$$S = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}\left(\frac{p(i,j)-\mu}{\sigma}\right)^3$$

Where, p(i, j) is the pixel value at point (i,j), m and σ are the mean and standard deviation respectively.

6. Kurtosis

Kurtosis, K measures the Peakness or flatness of a distribution relative to a normal distribution. The conventional definition of kurtosis is:

$$K = \left\{\frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}\left[\frac{p(i,j)-\mu}{\sigma}\right]^4\right\} - 3$$

Where, p(i,j) is the pixel value at point *(i,j),* m and σ are the Mean and Standard Deviation respectively. The -3 term makes the value zero for a normal distribution.

7. Root Mean Square (RMS)

The RMS *(Root Mean Square)* computes the RMS value of each row or column of the input, along vectors of a specified dimension of the input, or of the entire input. The RMS value of the j[th] column of an M-by-N input matrix u is given by below equation:

$$y = \sqrt{\frac{\sum_{i=1}^{M}|u_{ij}|^2}{M}}$$

8. Inverse Difference Moment (IDM)

It is a measure of image texture. IDM ranges from 0.0 for an image that is highly textured to 1.0 for an image that is untextured. The formula for finding the IDM is given in below equation:

$$H = \sum_{i,j}\frac{P(i,j)}{1+|i-j|}$$

9. Energy

Energy returns the sum of squared elements in the Grey Level Co-Occurrence Matrix (GLCM). Energy is also known as uniformity. The range of energy is [0 1]. Energy is 1 for a constant image. The formula for finding energy is given in below equation:

$$E = \sum_{i,j} P(i,j)^2$$

10. Contrast

Contrast returns a measure of the intensity contrast between a pixel and its neighbour over the whole image. The range of Contrast is [0 (size (GLCM, 1)-1) ^2]. Contrast is 0 for a constant image. Contrast is calculated by using the equation given below:

$$C = \sum_{i,j}|i-j|^2 P(i,j)$$

11. Correlation

Correlation returns a measure of how correlated a pixel is to its neighbor over the whole image. The range of correlation is [-1 1]. Correlation is 1 or -1 for a perfectly positively or negatively correlated image. Correlation is NaN (Not a Number) for a constant image. The below equation shows the calculation of Correlation.

$$Corr = \sum_{i,j}\frac{(i-\mu_i)(j-\mu_j)P(i,j)}{\sigma_i\sigma_j}$$

Where μ_i , μ_j , σ_i , and σ_j are the means and standard deviations of P_i and P_j, the partial probability density functions.

12. Homogeneity

Homogeneity returns a value that measures the closeness of the distribution of elements in the GLCM to the GLCM diagonal. The range of Homogeneity is [0 1]. Homogeneity is 1 for a diagonal GLCM. The Homogeneity is evaluated using the equation:

$$H = \sum_{i,j}\frac{P(i,j)}{1+|i-j|}$$

13. Variance

Variance is the square root of standard deviation. The formula for finding Variance is:

$$Var = \sqrt{SD}$$

Where SD is the Standard Deviation.

After extracting the features of segmented mass/tumor, then the dataset has to be constructed in the proper format, so that it can be given to any of the standard classifier tools.

Experimental Output

In Fig. 5, we can observe the values of the features extracted. Similarly the features have to be extracted for more number of images. The first image is the Mammogram input image and the second image is the tumor extracted by using the segmentation technique.

Future Enhancements

After calculating all the features for the set of pre diagnosed mammogram images, the feature dataset has to be constructed in the appropriate format, so that the classifier can understand. Some of the classifiers that can be used are LIBSVM, SVMLight, SVMtorch, ANN or any other classifiers. This phase is "Training Phase". Then in Testing Phase, for unknown (not diagnosed) mammogram images, features are determined and also feature dataset has to be constructed. The newly constructed feature dataset is given for the classifier which has been used in the training phase.

The classifier efficiency has to be determined, Receiver Operating Characteristics (ROC) plots has to be plotted. The classifier results of the unknown images/samples will be given for the radiologists for cross verification.

References

[1] Tabar L. and Dean P.B. (2003) *Gynaecol Obstet*, 82, 319-326.

[2] Giger M.L., Karssemeijer N. and Armato S.G. (2001) *IEEE Trans. on Med. Imaging*, 20, 1205-1208.

[3] Giger M.L. (2000) *Comput. Science Engineering*, 2, 39-45.

[4] Doi K., MacMahon H., Katsuragawa S., Nishikawa R.M. and Jiang Y. (1999) *Eur. J. Radiol*., 31, 97-109.

[5] Vyborny C.J., Giger M.L. and Nishikawa R.M. (2000) *Radiologic Clinics of North America*, 38, 725-740.

[6] Guido M. te Brake and Nico Karssemeijer (1999) *IEEE Transactions on Medical Imaging*, 18(7), 628-638

[7] Harvey J.E., Fajardo L.L., and Inis G.A. (1993) *AJR*, 161, 1167 -1172.

[8] Karssemeijer N.(1998) *Phys. Med. Biol.*, 43, 365-378.

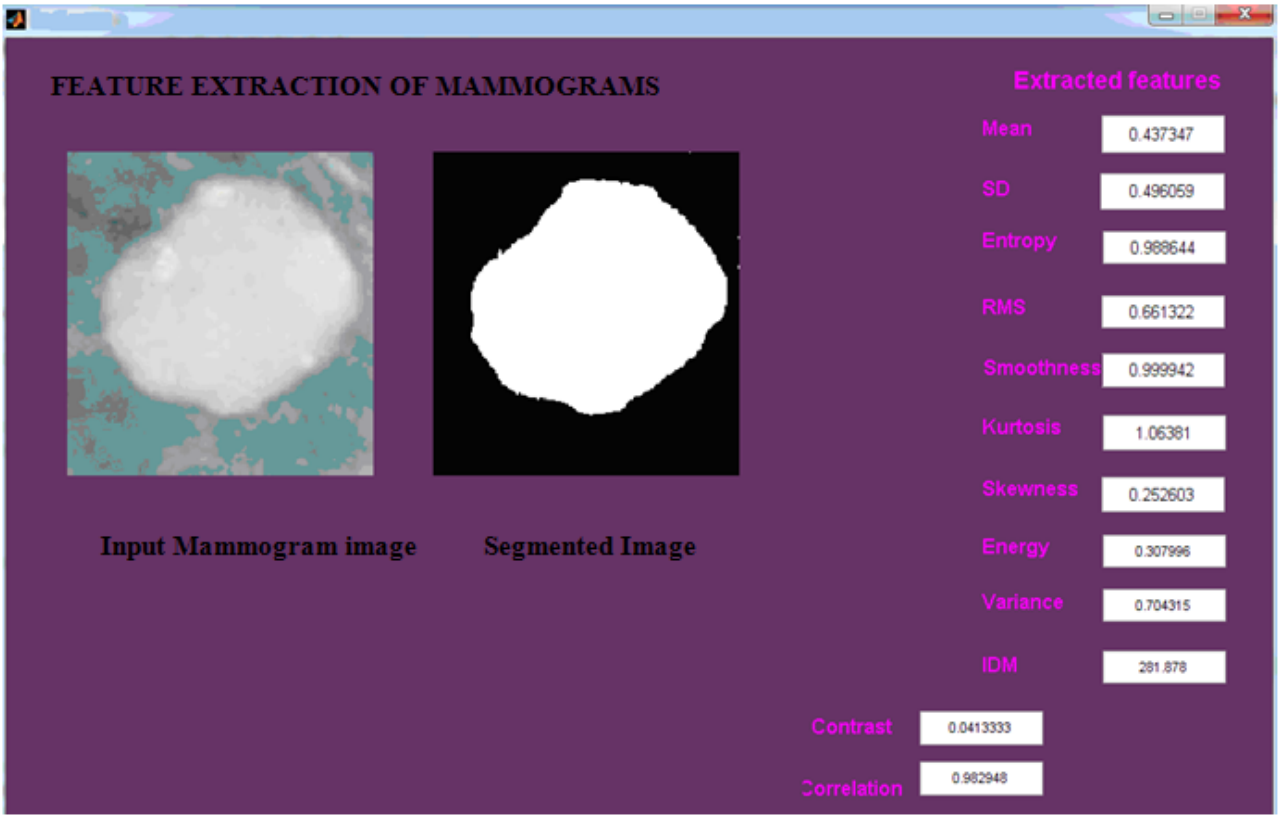

Fig. 5- Output Snapshot.