Feature Extraction and Classification of Mammographic Masses

Jignesh Panchal, Member, IEEE, Anuradha Agatheeswaran, Member, IEEE
Department of Electrical & Computer Engineering, Mississippi State University

***Abstract*-The aim of this project is to classify the mammographic masses as benign or malignant using texture and shape features. A set of 73 mammograms is used for the analysis, out of which 41 are benign and 32 are malignant. Manually segmented masses are obtained from the DDSM, USF database [2]. Texture and shape features are extracted from the manually segmented masses. Stepwise linear discriminant analysis is used to get the optimum set of features. Maximum-likelihood classifier with linear discriminant analysis (LDA) is used for the classification. The system is tested using leave-one-out test method and an overall accuracy of about 78 % is achieved.**

***Index Terms*- Gray-Tone Spatial-Dependence matrices, radial distance measure (RDM), feature extraction, and mammography.**

I. INTRODUCTION

Breast cancer ranks second among the cancer deaths in women [6]. Scientific studies show that earlier detection of cancer reduces the mortality rate. Mammography is a well known and one of the efficient methods to detect tumors. Computer aided systems are developed to aid the radiologists in detecting the tumor. This kind of analysis solves problems like high costs and need of highly skilled radiologists for mass screening. In recent years, digital mammography has proved to be a widely used image processing approach for the detection of breast cancer. Based on the texture and shape features, tumor is classified as benign and malignant.

As described in [3], normally mammograms have fibrous tissues, vessels, etc., while abnormal mammograms contain linear structures like spicules. Yajie Sun *et al*. discussed an algorithm to identify normal mammogram regions, which allow radiologists to concentrate more on suspicious areas [1]. Such an automated methods for diagnosis improves the performance by indicating suspicious areas. In [2] Bruce *et al.* talked about an automated mammographic mass shape classification system using 2-layer neural network classifier and wavelets decomposition.

The mammographic masses, considered for this study, are of two shapes: round and irregular. In shape analysis, the two-dimensional shape contour of each mass is mapped to a one-dimensional radial distance measure (RDM). The shape features (RDM features) are extracted from the manually segmented images. Since there is a chance of human error in marking the tumor boundary, the texture features are also extracted which is independent of shape features. Various methods for texture feature extraction have been proposed in the last couple of decades. In this paper, texture features are extracted from the gray-tone spatial-dependence matrices for the classification of tumors. In [4], Mudigonda *et al.* talked about the gradient based and texture based classification of mammographic masses using rubber band straightening technique (RBST) and gray-level co-occurrence matrices. They investigated the efficacy of textural information present in mass region and in the ribbon pixels around their boundaries.

In this project, an automated computer aided system is developed for the classification of breast tumors as benign or malignant based on shape features and texture features. The database for this system is obtained from the 'Digital data base for screening Mammography', University of South Florida, Tampa [2].

II. SYSTEM OVERVIEW

Figure 1 shows the general block diagram of entire system.

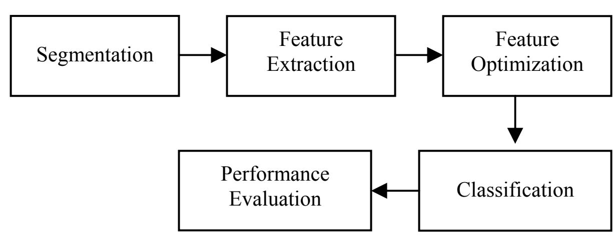


Figure 1. System Overview

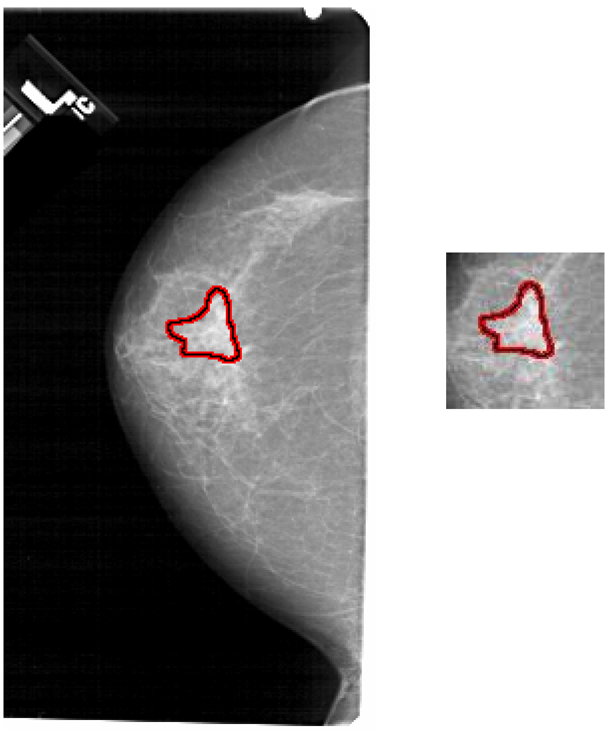
The tumor boundaries used in this project have been manually marked by expert radiologists. Textural and shape features are extracted for further analysis of the problem. Texture features like energy (angular second moment), contrast, correlation, inverse difference moment and entropy are extracted using gray-level co-occurrence matrices and shape features like mean, variance, and zero-cross are extracted from RDM. Features are optimized using stepwise linear discriminant analysis. Maximum-likelihood classifier

Anuradha Agatheeswaran is with the Electrical Engineering Department, Mississippi State University (e-mail: aa130@ece.msstate.edu).
Jignesh Panchal is with the Electrical Engineering Department, Mississippi State University (e-mail: jsp71@ece.msstate.edu).

with LDA is used and system is tested using leave-one-out criteria. The following sections explain each block, as shown in figure 1, in detail.

III. Mammogram Database

In this project, a set of 73 manually segmented mammograms is analyzed, out of which 41 are benign and 32 are malignant. The software "heathusf v1.1.0" [5] is used to convert the compressed mammograms to 8-bit/pixels. The entire mammograms are not used for the analysis; instead the region of interest (ROI) is cropped to a size of 1024 x 1024 pixels with the tumor at center. The manually extracted mass borders are assumed to be correct. Using this method, a set of ROIs is generated, which are used as the dataset for entire system analysis.



(a) (b)

Figure 2. (a) Entire mammogram with boundary marked by radiologists
(b) Cropped ROI used for the analysis

IV. Feature Extraction

The shape features are based on radial distance measure (RDM) and texture features are based on the gray-tone spatial-dependence matrices. Three shape feature and five texture features are selected. Age of the patient is added as the ninth feature.

A. Shape features

The radial distance measure is the Euclidean distance calculated from center of the tumor to the boundary pixels and normalized by dividing with the maximum length [8]. The RDM plots for one of the benign and malignant cases are shown in figure 3 and figure 4 respectively. The features extracted from the RDM are mean, variance and zero crossing. The mean and variance are calculated using

$$d_{avg}=\frac{1}{N}\sum_{i=1}^{N}d(i) \qquad \sigma^2=\frac{1}{N}\sum_{i=1}^{N}\left(d(i)-d_{avg}\right)^2$$

where $d(i)$ is the normalized length, d_{avg} is the mean distance and σ^2 is the variance[9]. The mean and variance measure the changes in the boundary. The zero-crossing count is the measure of the roughness of the boundary and is the number of times the radial distance crosses the mean distance [9].

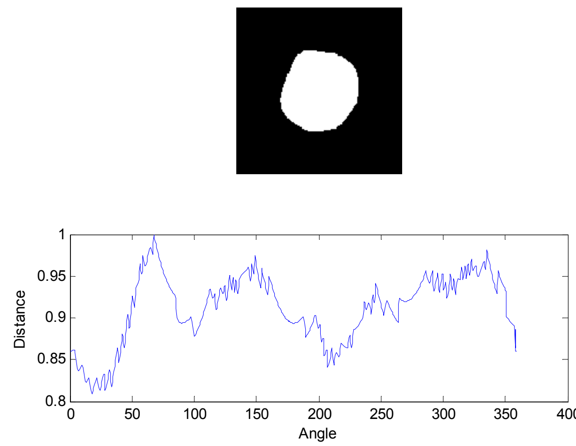


Figure 3. Template and radial distance measure for benign case

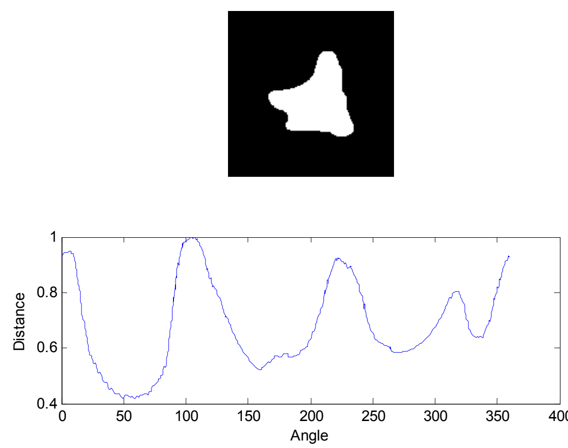


Figure 4. Template and radial distance measure for malignant case

B. Texture features

Texture features contain information about the spatial distribution. The key factor in texture analysis is extraction of texture features. The spatial gray level matrices are based on the estimation of the texture-context information contained in the overall or the average spatial relationship of the gray levels [7]. The gray-tone spatial-dependence frequencies are a function of the angular relationship and distance between neighboring resolution cells [7]. A distance of 1 pixel and four directions (0°, 45°, 90°, 135°) are considered in this paper. The gray-tone spatial-dependence matrix is normalized by dividing the maximum possible resolution cells for a given direction [7]. The sum of texture features in all directions is considered, so this reduces from 20 texture features to 5 texture features. Some of the texture features used are [7],[4]

- *Energy* (angular second moment) provides the information regarding the uniformity of the region. The energy will be low if all the elements in the matrix are equal.

$$f_1 = \sum_i \sum_j \{p(i,j)\}^2$$

- *Contrast* (difference moment) is a measure of the amount of local variations

$$f_2 = \sum_{n=0}^{N_g-1} n^2 \left\{ \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i,j) \right\}$$

- *Correlation* is the measure of gray tone linear dependence

$$f_3 = \frac{\sum_i \sum_j (ij) p(i,j) - \mu_x \mu_y}{\sigma_x \sigma_y}$$

- *Inertia* is the measure of degree of fluctuations of the image intensity

$$f_4 = \sum_i \sum_j (i-j)^2 p(i,j)$$

- *Homogeneity*

$$f_5 = \sum_i \sum_j \frac{1}{1+(i-j)^2} p(i,j)$$

N_g is the number of gray levels , $p(i,j) = P(i,j)/R$, where $P(i,j)$ is the unnormalized gray-tone spatial dependence frequencies and R is the normalizing constant. μ_x, μ_y, σ_x and σ_y are the mean and standard deviation of the marginal probability matrices p_x and p_y. p_x and p_y are define as [7]

$$p_x = \sum_{j=1}^{N_g} P(i,j)/R \quad \text{and} \quad p_y = \sum_{i=1}^{N_g} P(i,j)/R$$

V. CLASSIFICATION

Stepwise linear discriminate analysis is used to get the best combination of features. Stepwise LDA includes forward selection and backward rejection. The flowcharts (figure 5) describe the forward selection and the backward rejection. The maximum likelihood classifier is used to evaluate the performance of a feature. A set of 32 cases is jack-knifed and used to train and test.

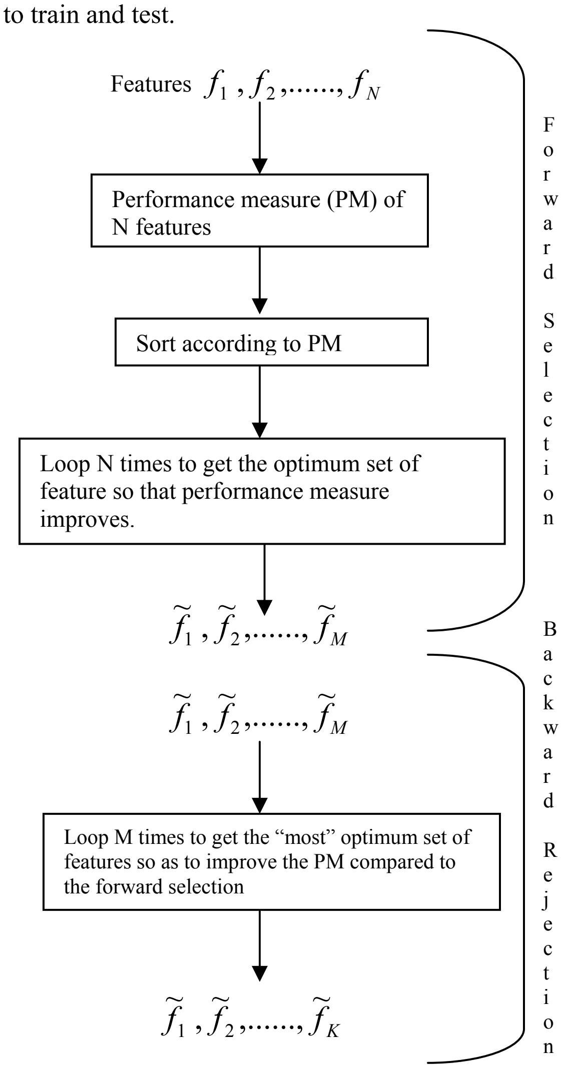


Figure 5. Flow diagram for stepwise LDA [10]

Stepwise LDA helps to get the "best" optimum set of combination of features. In the forward selection, features are first sorted in descending order of the performance measure (PM). The first feature is included in the feature vector. The second feature is added to the feature vector and the PM of combination of these features i.e. $PM_1 > PM_2$, second feature is retained else eliminated. The third feature is added to the feature vector and if $PM_2 > PM_3$, third feature is retained else eliminated [10]. Optimum set of features is obtained at the end of forward selection. The output of forward selection is input to the backward rejection. The first feature of the optimum set is removed. If PM of the feature vector excluding the first feature is better than the total feature vector then the first feature is eliminated. This procedure is repeated for all the features resulting from the forward selection. Thus the "most" optimum set of features are obtained at the end of backward rejection.

The "most" optimum set of features are used to classify the mass as benign or malignant. A maximum likelihood classifier with LDA is used for classification with leave-one out testing method. The performance measure uses maximum likelihood technique so that the accuracies match that of the classifier.

VI. Results and Discussions

The performance measure of each of the features is tabulated in table 1. The texture and shape features are used to classify the masses individually. Table 2 shows the results for accuracies for texture and shape feature individually. The classification accuracies after stepwise LDA are shown in table 3. The features got selected after stepwise LDA are age, entropy, RDM variance, RDM mean and inertia. It can be seen that there is an improvement in the accuracy when both texture, shape and age features are used instead of only texture or shape features.

TABLE 1: ACCURACIES OF INDIVIDUAL FEATURES

	Feature	Accuracy
Texture	Energy	0.5625
	Inertia	0.40625
	Entropy	0.65625
	Homogeneity	0.46875
	Correlation	0.46875
Shape	RDM mean	0.53125
	RDM variance	0.46875
	Zero-crossings	0.375
	Age	0.65625

TABLE 2 (A): CONFUSION MATRIX FOR TEXTURE FEATURES

	Benign	Malignant	
Benign	24	17	0.5854
Malignant	19	13	0.4063
	0.5581	0.4333	**0.5068**

TABLE 2 (B): CONFUSION MATRIX FOR SHAPE FEATURES

	Benign	Malignant	
Benign	38	3	0.9268
Malignant	27	5	0.1563
	0.5846	0.6250	**0.5890**

TABLE 3: CONFUSION MATRIX FOR THE OPTIMUM SET OF FEATURES AFTER PERFORMING STEPWISE LDA

	Benign	Malignant	
Benign	34	7	0.8293
Malignant	9	23	0.7188
	0.7907	0.7667	**0.7808**

VII. Conclusions

In this study, 73 mammograms are analyzed for texture and shape features. The overall accuracy obtained using only texture analysis was around 50%, and by using only shape analysis was around 59% whereas the accuracies were increased with optimum combination of texture, shape and age features. The overall accuracy obtained with this combination is 78%.

Future Work:

The malignant masses segmented by the radiologists are not very accurate. Therefore, the shape features like zero-crossing didn't give a good result. Thus the more accurate segmentation method can be implemented. Texture features can be extracted from the RBST. Different algorithms for texture feature like gray-level run length method, gray level difference method can be implemented [11].

VIII. Reference

[1] "Normal mammogram classification based on regional analysis" -Yajie Sun; Babbs, C.F.; Delp, E.J.; Circuits and Systems, 2002. MWSCAS-2002. The 2002 45th Midwest Symposium on, Volume: 2 , 4-7 Aug. 2002

[2] http://marathon.csee.usf.edu/Mammography/Database.html

[3] "Classification of Linear Structures in Mammographic Images - Reyer Zwiggelaar and Caroline R.M. Boggis, Division of Computer Science, University of Portsmouth, Greater Manchester Breast Screening Service, Withington Hospital, Manchester

[4] "Gradient and texture analysis for the classification of Mammographic masses" *Mudigonda, N.R.; Rangayyan, R.; Desautels, J.E.L.;* Medical Imaging, IEEE Transactions on, Volume: 19, Issue: 10, Oct. 2000 Pages: 1032 – 1043

[5] http://marathon.csee.usf.edu/Mammography/software/heat heathusf_v1.1.0.html

[6] www.**cancer**.org/downloads/STT/CAFF2003PWSecured.pdf

[7] "Texture Features for image Classification" *Haralick , R.M; Shanugam, k; Dinstein, I;* Systems,Man and Cybernetics,IEEE transactions on, Vol.SMC-3,No. 6 Nov. 1973 Pages 610 – 621

[8] "Classifying Mammograhic Lesions Using Computerized Image Analysis" *Kilday, J; Palmieri, F; Fox, M.D;* Medical Imaging, IEEE Transactions on, Volume: 12, No.4, 1993, Pages: 664 – 669

[9] "Classifying Mammographic Mass Shapes Using the wavelet transform Modulus-Maxima Method" *Bruce, L.M; Adhami, R.R;* Medical Imaging, IEEE Transactions on, Volume: 18,No.12,Dec 1999, Pages: 1170 – 1177

[10] **"**Discrimination of subtly different vegetative species via hyperspectral data" *Mathur, A.; Bruce, L.M.; Byrd, J;* Geoscience and Remote Sensing Symposium, 2002. IGARSS '02. 2002 IEEE International , Volume: 2 , 2002 Page(s): 805 –808

[11] "A Theoretical Comparison of Texture Algorithms " *Conners, R.W; Harlow, C.A;* Pattern Analysis and Machine Intelligence, IEEE Transactions on, Vol: PAMI-2, No. 3, May 1980, Pages 204 - 222