Using Haralick Features for the Distance Measure Classification of Digital Mammograms

B. Kishore
Asst. Professor – Sr. Scale
CSE Dept., MIT, Manipal

R. Vijaya Arjunan, Ph.D.
Asst. Professor – Sr. Scale
CSE Dept., MIT, Manipal

Rupsa Saha
Student
CSE Dept., MIT, Manipal

Siva Selvan
Asst. Professor
CSE Dept., MIT, Manipal

ABSTRACT
Texture analysis is one of the primary ways of extracting relevant information from digital images. Analysis of digital mammograms is essential in distinguishing between normal tissue and tissues that are showing early signs of breast cancer. In this paper, we compute certain Haralick texture features (Angular Second Moment, Contrast, Correlation and Entropy) and compare the performance of simple distance-measure classifications with each of these features, as well as the mean of all four. The correlation feature and the mean of all four features shows better accuracy when applied on digital mammograms to classify them into normal tissues and cancerous tissues.

General Terms
Analysis, extraction, feature, classification

Keywords
Mammography, texture, normal, cancerous

1. INTRODUCTION
In recent times, advances in information technology and computer science have brought about many advancements in the health care sector. The field of medical imaging has undergone a major technological overhaul due to digital imaging advances over the last forty years, and digital imaging has almost completely replaced film radiographic techniques. Digital records of patients provide easier storage, faster transmission and retrieval as well as a higher accuracy of analysis, all of which have revolutionized modern healthcare. Integrated "teleradiology" services can provide high-performance computing facilities [4], which can execute computationally intensive image analysis and visualization tasks. This is used by medical professionals for faster and more accurate identification of various diseases and conditions. Ongoing research consists of extensive work on building systems which can detect abnormalities from medical images with reduced manual intervention. This can help cut down on the time taken for analysis, making premier healthcare and screening facilities available to a wider section of society, as well as addressing the shortage of qualified medical professionals. It should be mentioned here that fully automated abnormality detection/identification is still not possible, as of now, due to the low margin of error that can be afforded in the medical field.

Breast cancer one of the most common malignancies in women. According to current medical data, it accounts for almost 22.9% of all cancers in women [6]. Screening is the most effective way of diagnosing breast cancer, with physical examination and mammography being the two most popular methods for the same. Early detection and diagnosis of breast cancer is essential to increase the chances of survival and complete recovery. Screening mammography, which is currently one of the best available radiological technique for detection of breast cancer [1], is an x-ray examination of the breasts of (usually, an asymptomatic) woman. Screening mammography enables detection of early signs of breast cancer such as masses, calcifications and architectural distortion and bilateral asymmetry. On the other hand, diagnostic mammography examination is performed for symptomatic women who have an abnormality found during screening mammography [5]. The breast image is captured using a special electronic x-ray detector which converts the image into a digital mammogram for viewing, storing and analysis. Each breast is imaged separately in craniocaudal (CC) view and mediolateral-oblique (MLO) view.

Any kind of digital image analysis requires the extraction of features by transforming the data in the high-dimensional space to a space of fewer dimensions. A transformed image is represented by a set of features, known as a feature vector. Texture is an important component for analysis of images and identification of regions of interest. Texture analysis in medical images involves techniques to evaluate the position and intensity of pixels, and their grey-level intensity which gives information about the objects found in those images [8].

One of the most popular approaches to texture analysis is based on the co-occurrence matrix obtained from images, proposed by Robert M. Haralick in 1973 [2], which forms the basis of this paper. We have studied the implementation of certain Haralick features on a collection of digital mammography images, obtained from the MIAS database of mammograms. The results obtained are discussed further in this work.

Classification of images based on the information obtained via analysis is an important aspect of building systems that can detect abnormalities or deviations automatically. Classification of images set can be done based on the feature vectors obtained from the entire image set. In this study, we make use of MATLAB classification feature, using different types of distance classification methods, like linear, quadratic, mahalanobis, diaglinear etc.

2. LITERATURE SURVEY
Features are extracted from images by considering groups of pixels, the information provided by those groups and/or known relevant background information. Feature extraction is essential for classification and recognition of images. A number of automated feature extraction techniques are available to modern digital image processing. Major among these are adaptive linear filters, clustering methods, and artificial neural networks. [10] Commonly extracted features

are shape, size, perimeter, intensity, spatial orientation, Haralick textures, colour, Zernike moments, etc. [11]

Humans perceive and differentiate between images using three major parameters: spectrum, texture and context. In the analysis of black and white photographs, tone represents the varying gray levels in resolution cells, while the statistical distribution of the gray levels is interpreted as texture.[12] Tone and texture form an intrinsic part of any image, and they are not independent concepts in themselves, though one can get precedence over the other according to the nature of the image. The relation between the two can be expressed in simple terms as tone is dominant when the sample under consideration shows only a minor variation of gray levels over a small area, whereas a wide range of variation of the same indicate the dominance of texture.[13]

A primary assumption in Haralick's paper is, textural information of any image is contained in the average relation of the gray tones of the image with each other in the spatial domain. The procedure is based a set of gray-tone spatial-dependence probability distribution matrices (also termed as Gray-Level Co-occurrence Matrices or GLCM), computed for various angular moments and distances, from each of which a total of thirteen features can be extracted.[2] The GLCM is now one of the most widely used methods for texture classification. Each cell (i,j) in a GLCM corresponds to the number of occurrences of the pair of gray levels i and j, which are a distance of d from each other in image. [14] These features provide information about image texture in terms of contrast, homogeneity, linear variation of gray tone, boundaries etc.

Co-occurrence Matrix: A co-occurrence matrix, P, is defined to describe the patterns of neighboring pixels in an image occurring at a given distance 'd'. 4 co-occurrence matrices, each at a different angular orientation, are used for the calculation of the texture features. [15] A co-occurrence matrix that describes pixels that are adjacent to one another at angle 0^o (i.e. horizontally) is termed as p_0. Similarly, there are co-occurrence matrices corresponding to the vertical direction (90^o) and both diagonals (45^o and 135^o). These matrices are called p_{90}, p_{45} and p_{135} respectively [3]. The creation of co-occurrence matrices with d=1 is given below.

$$x = \begin{pmatrix} 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 2 & 2 & 2 \\ 2 & 2 & 3 & 3 \end{pmatrix} \quad \text{…….. (a)}$$

$$p_0 = \begin{pmatrix} 4 & 2 & 1 & 0 \\ 2 & 4 & 0 & 0 \\ 1 & 0 & 6 & 1 \\ 0 & 0 & 1 & 2 \end{pmatrix} \quad \text{…….. (b)}$$

$$p_{45} = \begin{pmatrix} 4 & 1 & 0 & 0 \\ 1 & 2 & 2 & 0 \\ 0 & 2 & 4 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix} \quad \text{…….. (c)}$$

$$p_{90} = \begin{pmatrix} 6 & 0 & 2 & 0 \\ 0 & 4 & 2 & 0 \\ 2 & 2 & 2 & 2 \\ 0 & 0 & 2 & 0 \end{pmatrix} \quad \text{…….. (d)}$$

$$p_{135} = \begin{pmatrix} 2 & 1 & 3 & 0 \\ 1 & 2 & 1 & 0 \\ 3 & 1 & 0 & 2 \\ 0 & 0 & 2 & 0 \end{pmatrix} \quad \text{…….. (e)}$$

Fig 1: (a) Example matrix. (b) Matrix p_0 (c) Matrix p_{45} (d) Matrix p_{90} (e) Matrix p_{135}

There are 4 pairs of (0,0) in angular 0, thus $p_0(0,0)=4$, there are 2 pairs of (0,1), thus $p_0(0,1)=2$. Similarly all the four matrices are computed. The thirteen features proposed by Haralick are given in Appendix I.

3. METHODOLOGY

In this paper, we choose four out of the thirteen features proposed by Haralick in order to make the task computationally effective. These features are used to classify mammography images taken from the Mini-Mias Database into two categories "Normal" and "Cancerous" [9].

We consider the features Angular Second Moment (f1), Contrast (f2), Correlation (f3) and Entropy (f9).

Angular Second Moment

$$f_1 = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i,j)^2$$

Contrast

$$f_2 = \sum_{k=0}^{N_g - 1} k^2 p_{x-y}(k)$$

Correlation

$$f_3 = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} (ij) p(i,j) - \mu_x \mu_y}{\sigma_x \sigma_y}$$

Entropy

$$f_9 = -\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i,j) \log\left(p(i,j)\right)$$

The work involves first calculating the features of the training set and using it to train the classification module.

The function "imread" is used to convert image data into a 256 gray-level matrix in MATLAB.

For fast calculation of spatial-dependence matrices, the gray tone data has to be quantized with respect to either 2 or 4 or 8 gray levels as shown in the following table. This produces a compressed image with less redundant information than the original.

Table 1. Compression of image data

Original	2 gray tone	4 gray tone	8 gray tone
0~31	0	0	0
32~63			1
64~95		1	2
96~127			3
128~159	1	2	4
160~191			5
192~223		3	6
224~255			7

In the next step, the four two-dimensional GLCM matrices with respect to angles 0, 45, 90 and 135 degrees are calculated. These four matrices are combined together to form one single three dimensional data cube, which is passed to each of the functions that compute the feature vectors mentioned above. Each vector is a 4X1 matrix, with values corresponding to the 4 angular moments. For supervised learning, 10 and 20 images of each category (normal and cancerous) are used for training data in our experiment. The mean of all directional data from each feature vector is stored as the training data for classification. The training data and the samples to be classified are then passed to the MATLAB function classify [7]. This function is implemented as follows: class = classify (sample, training, group), where each row of the data in sample is classified into one of the groups in training. The various types of classification that can be performed using this simple function are [16]:

- Linear - Fits a multivariate normal density to each group, with a pooled estimate of covariance. This is the default.
- Diaglinear - Similar to linear, but with a diagonal covariance matrix estimate (naive Bayes classifiers).
- Quadratic - Fits multivariate normal densities with covariance estimates stratified by group.
- Diagquadratic - Similar to quadratic, but with a diagonal covariance matrix estimate (naive Bayes classifiers).
- Mahalanobis - Uses Mahalanobis distances with stratified covariance estimates.

4. RESULTS AND ANALYSIS

The analyses of the results obtained are detailed below. Analysis is done on the basis of the mean values of each of the feature vectors obtained.

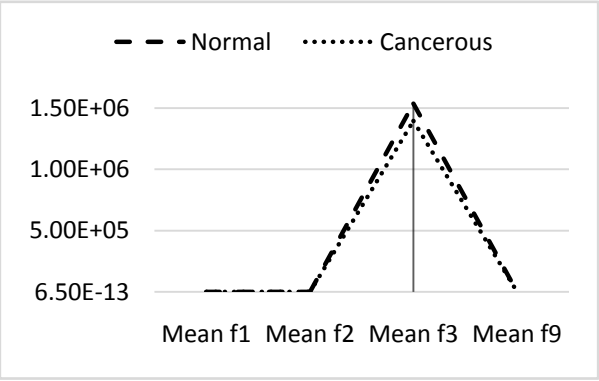

Fig 2: Comparison of individual mean values of the four selected features

Fig 2 is a graphical representation of the mean values of f1, f2, f3 and f9 for images of normal tissue and for images showing early signs of breast cancer. For f1, f2 and f9 the mean values almost coincide for normal and cancerous data, providing very little separation between the two classes of images.

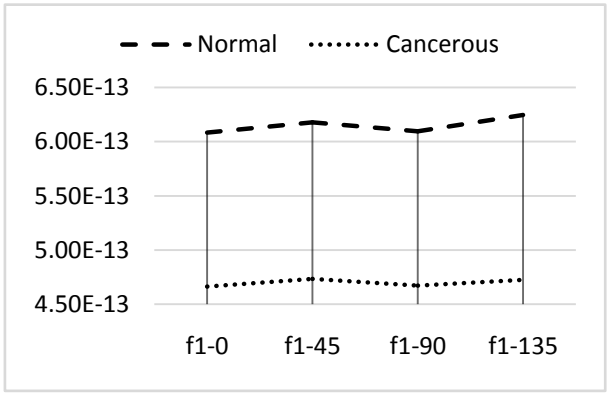

Fig 3: Differences in mean values of f1

Fig 3 shows the distribution of values of Feature 1, i.e. Angular Second Moment, over the angles 0, 45 90 and 135 degrees, for both images of normal tissue and for images showing early signs of breast cancer.

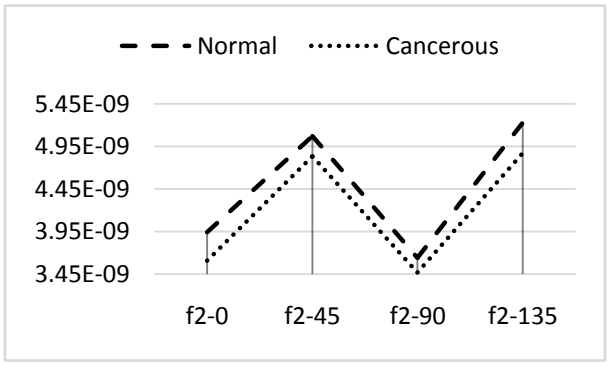

Fig 4: Differences in mean values of f2

Fig 4 represents the values obtained for Feature 2, i.e. Contrast, over the angles 0, 45 90 and 135 degrees, for images of normal tissue as well as for images showing early signs of breast cancer.

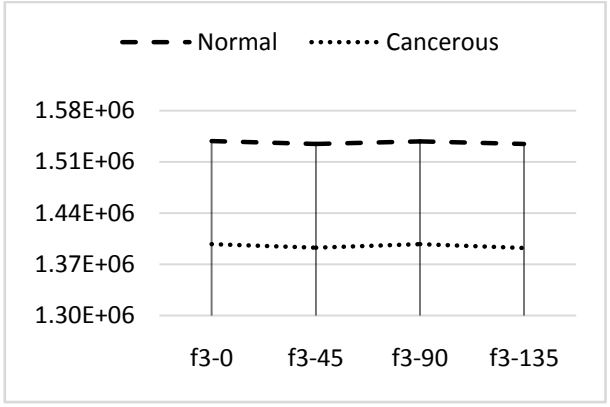

Fig 5: Differences in mean values of f3

Fig 5 shows the distribution of values of Feature 3, i.e. Correlation, over the four angles, for separate sets containing images of normal tissue and those showing early signs of breast cancer.

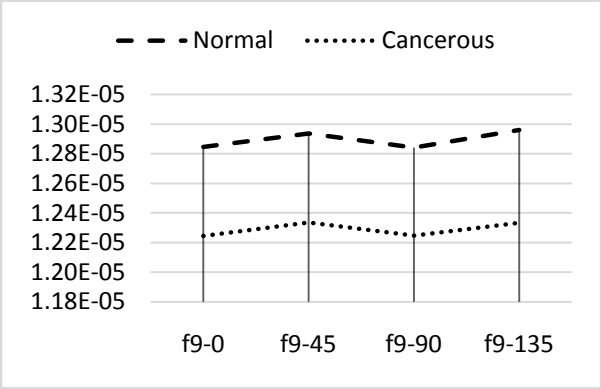

Fig 6: Differences in mean values of f9

In a similar manner, Fig 6 shows the values computed for Feature 9, i.e. Entropy, for each of the four angles 0,45, 90and 135 degrees, for both sets of images of normal tissue and for images showing early signs of breast cancer.

The average difference between the feature values of normal tissue images and cancerous tissue images are 1.45E-13 for f1, 2.75E-10 for f2, 1.41E+05 for f3, 6.05E-07 for f9 and 3.53E+04 for the mean of all four features. This is demonstrated in Fig 7 given below.

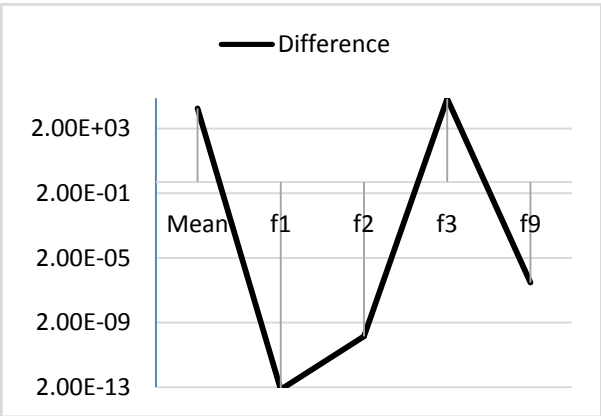

Fig 7: Average difference between feature values of normal tissue images and cancerous tissue images

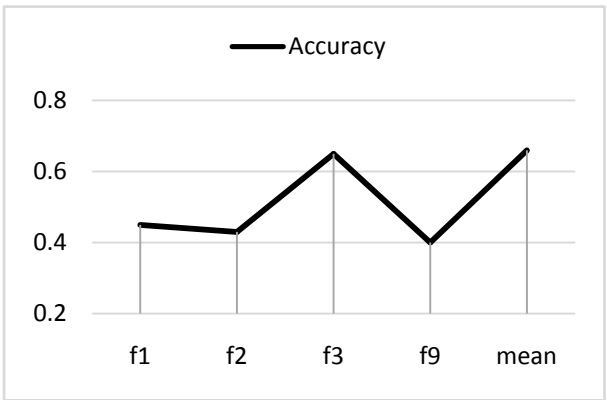

Fig 8: Accuracy of classification of images into normal and cancerous

Using simple distance measure classification on all five sets of features obtained above, we see that f3 (i.e. Correlation) and the mean give the best results for classification. This can be understood intuitively since the other features show very little difference in feature vector values for the different classes, and is proved experimentally. The accuracy of classification of images into "normal" and "cancerous", as obtained from MATLAB using the function classify(), is shown in Fig 8 given above.

5. CONCLUSION

As discussed in the results and analysis section the correlation feature and the combined mean of all four features shows better accuracy when applied on digital mammograms to classify them into normal tissues and cancerous tissues. Statistical measures of other Haralick texture features may also be tested on the similar data set to identify a sub set of features that gives maximum accuracy for classification. Various combinations like mean, median, standard deviation etc., of different subsets of Haralick features may produce varying degrees of accuracy for classification, based on the type of images and the type of features under consideration. Another useful addition to this research would be to study the effect of parallel processing in this context to reduce the computation time taken.

6. REFERENCES

[1] Timo Ojala, Matti Pietikainen and David Harwood, A Comparative Study Of Texture Measures With Classification Based On Feature Distributions, Pattern recognition 29.1 (1996): 51-59

[2] Robert M. Haralick, K. Shanmugam and ITS'HAK Dinstein. Textural features for Image Classification, IEEE Transactions on Systems, Man, and Cybernetics, 1973, SMC-3 (6): 610–621

[3] Eizan Miyamotol and Thomas Merryman Jr, Fast Calculation of Haralick Texture Features, Technical Report, Carnegie Mellon University, found at: www.ece.cmu.edu/pueschel/teaching/18-799B-CMU-spring05/ material/eizan-tad.pdf

[4] M. Zikos, E. Kaldoudi, S. C. Orphanoudakis. DIPE: A Distributed Environment for Medical Image Processing In Proceedings of MIE'97 Medical Informatics Europe, 14th International Congress, Porto Carras, Greece, May 25-29, 1997

[5] Jelena Bozek, Mario Mustra, Kresimir Delac, and Mislav Grgic. A Survey of Image Processing Algorithms in Digital Mammography.

[6] "World Cancer Report". International Agency for Research on Cancer, 2008.

[7] Mathworks R2014b Online Documentation Online [http://www.mathworks.in/help/stats/classify.html].

[8] G. Castellano, L. Bonilha, L.M. Li, F. Cendes Texture analysis of medical images, Neuroimage Laboratory, Faculty of Medical Sciences, State University of Campinas, Brazil. Clinical Radiology (2004) 59, 1061–1069.

[9] J Suckling et al, The Mammographic Image Analysis Society Digital Mammogram Database Exerpta Medica. International Congress Series, 1994 1069 pp 375-378.

[10] Michael Brennan, Comparison of automated feature extraction methods for image based screening of cancer cells, Uppsala Universitat, January 2012.

[11] Steven A. Hane, High Content Screening: Science, Techniques and Applications, Wiley Publishers,2008, pp.66

[12] Sharma et al., Mathematical Modeling In Geographical Information System (gis) & Gps An Overview, Concept Publishing Company, 2006, pp 56.

[13] Robert M. Haralick, Statistical and Structural Approaches to Texture, Proceedings of the IEEE, Vol. 67, No. 5, May 1979

[14] M. Partio et al., Rock texture retrieval using gray level co-occurrence matrix, 5th Nordic Signal Processing Symp., 2002

[15] V. Bino Sebastian at al., Gray level co-occurrence matrices: generalisation and some new features, International Journal of Computer Science Engineering and Information Technology, 2 (2012), pp. 151–157

[16] Classification toolbox for MATLAB, Milano Chemometrics and QSAR Research Group, found at: http://michem.disat.unimib.it/chm/download/softwares/help_classification/theory.htm

Appendix – I

Notation:

N_g : Number of distinct gray levels in quantized image

$$p_x(i) = \sum_{j=1}^{N_g} p(i,j)$$

$$p_y(j) = \sum_{i=1}^{N_g} p(i,j)$$

$$p_{x-y}(k) = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p(i,j),\ |i-j| = k \text{ and } k = 0,1,\ldots,N_g - 1$$

$$p_{x+y}(k) = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p(i,j),\ i+j = k \text{ and } k = 2,3,\ldots,2N_g$$

a) Angular Second Moment

$$f_1 = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p(i,j)^2$$

b) Contrast

$$f_2 = \sum_{k=0}^{N_g-1} k^2 p_{x-y}(k)$$

c) Correlation

$$f_3 = \frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_g}(ij)p(i,j) - \mu_x\mu_y}{\sigma_x\sigma_y}$$

Where μ_x, μ_y, σ_x, σ_y are mean of x, y and standard deviation of x, y respectively.

d) Sum of Squares: Variance

$$f_4 = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g}(i-\mu)^2 p(i,j)$$

e) Inverse Difference Moment

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

f) Sum Average

$$f_6 = \sum_{i=2}^{2N_g} i p_{x+y}(i)$$

g) Sum Variance

$$f_7 = \sum_{i=2}^{2N_g}(i-f_8)^2 p_{x+y}(i)$$

h) Sum Entropy

$$f_8 = -\sum_{i=2}^{2N_g} p_{x+y}(i)\log\left(p_{x+y}(i)\right)$$

i) Entropy

$$f_9 = -\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p(i,j)\log\left(p(i,j)\right)$$

j) Difference Variance

$$f_{10} = \text{variance of } p_{x-y}$$

k) Difference Entropy

$$f_{11} = -\sum_{i=0}^{N_g-1} p_{x-y}(i)\log\left(p_{x-y}(i)\right)$$

l) Information measures of correlation

$$f_{12} = \frac{f_9 - HXY1}{\max(HX, HY)}$$

$$f_{13} = \sqrt{1-\exp^{-2(HXY2-f_9)}}$$

Where HX and HY are entropies of p_x and p_y.

$$HXY1 = -\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p(i,j)\log\left(p_x(i)p_y(j)\right)$$

$$HXY2 = -\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p_x(i)p_y(j)\log\left(p_x(i)p_y(j)\right)$$