International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 3 Issue 9, September- 2014

A Hybrid Content Based Image Retrieval for Classification of Mammograms

I. Naga Padmaja
M.TechScholar
Department of Computer Science & Engineering
VVIT, Nambur (V), Guntur (Dist.), India

T. Sudhir
Associate Professor
Department of Computer Science & Engineering
VVIT, Nambur (V), Guntur (Dist.), India

Dr. E. Srinivasa Reddy
Professor
ANU College of Engineering & Technology
Acharya Nagarjuna University, Guntur

***Abstract:* -Mammogram, a breast X-ray is the most effective, low cost, and reliable method for early detection of breast cancer. Cancer can be detected by classifying mammogram image into normal, benign and malignant class. This paper proposes an approach to develop a computer-aided classification system for cancer detection using Gray Level Run Length Matrix (GLRLM). The proposed method hybridizes texture feature extraction with a meta-heuristic, Genetic Algorithm, in the selection of optimal features for the mammogram classification. This technique provides more accuracy by data reduction in terms of feature selection and reduced time to classify the image.**

Keywords: -Mammogram, Texture, Gray level run length Matrix, Genetic Algorithm.

I. Introduction

Breast cancer is the second leading cause of cancer deaths among women. Early detection of the cancer allows treatment that could lead to high survival rate. Mammography is the process of using low-energy X-rays to examine the human breast and is used as a diagnostic and a screening tool. The goal of mammography is the early detection of breast cancer, typically through detection of characteristic masses and/or microcalcifications. However, 10–30 % of breast cancers are missed at mammography [1].

Mining information and knowledge from large database is became a key research topic in database system. In machine learning, the researchers are using data mining algorithms for the purpose of image learning[2-8].

Medical image mining is a process of extracting knowledge and providing scientific decision making for the diagnosis and treatment planning. Different methods of data mining have been used to detect and classify anomalies in mammogram images such as wavelets [9,10] and statistical methods. Most of them have used features extraction using image processing techniques[5]. Some other methods are based on fuzzy theory [1] and neural networks [11]. Haralick et al. [15] have used 14 different features from the co-occurrence matrix. Aswini Kumar Mohanty et al. efficiently extracted the GLCM features using Genetic algorithm [12]. In another work, Aswini Kumar Mohantyetal.classified mammograms using GLRL matrix [13]. GLRLM features are used for the purpose of classification of prostate cancer by ManavalanRadhakrishnanand Thangavel Kuttiannan2 [14]. Now in this paper we are proposing a hybrid algorithm for the selection of optimal feature of GLRLM. The proposed method extracts seven features using GLRLM from each input image and selects optimal features using GA that minimizes the classification error rate. Experiments are conducted on various mammogram images of MIAS database. The Experimental results have shown the improved accuracy of the proposed algorithm.Section II explains the methodology for mammogram retrieval using the proposed gray level statistical matrix. Section III presents the experimental results and discussions. Finally, Section IV gives the conclusion.

II. Methodology

Classification of mammograms using the proposed Genetic algorithm consists of feature extraction and image retrieval. During the first stage, in the pre-processing step the regions of interest (ROIs) of the database images are normalized to zero mean and unit variance [16]. From the pre-processed images, the gray level run length matrix (GLRLM) is generated in order to estimate the texture features. Finally, the performance measures are calculated using SVM Classification in order to analyse the effectiveness of the proposed method towards mammogram retrieval.

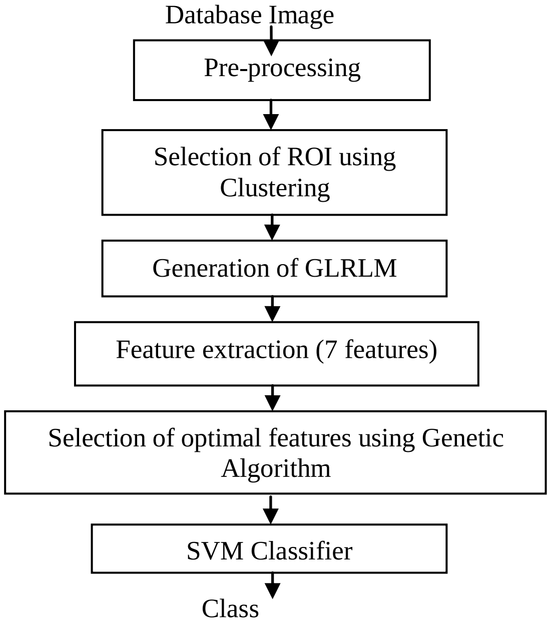

Database Image
Pre-processing
Selection of ROI using Clustering
Generation of GLRLM
Feature extraction (7 features)
Selection of optimal features using Genetic Algorithm
SVM Classifier
Class

Fig. 1 Overview of mammogram retrieval using the proposed approach

A. *Gray- level co-occurrence Matrix (GLCM)*

In Image processing a region texture can be described using three approaches. They are statistical, structural and spectral. Properties of textures are computed based on spatial distribution of gray levels in statistical approach. The statistical methods can be divided into three types. They are first order, second order, high order statistics. First order statistics considers only individual pixel intensity values, avoiding the spatial communication between the pixel values for estimating the texture properties. Second order statistics considers the occurrence of two or more pixel values at specific locations relative to each other for estimating the texture properties.

GLCM is second order statistics, which is used to describe the occurrence of combinations of gray level values in an input image with a specified distance and direction. The "Fig. 2(a)" represents example of an image with 4 gray levels. The "Fig 2(b)" represents the GLCM of image with a distance 1 unit and direction o◦. Different GLCMs can be generated with four directions, 0^0, 45^0,90^0,135^0. These directions are shown in "Fig 2(c)".

B. *Gray Level Run-Length Matrix(GLRLM)*

Like GLCM, GLRLM also used to extract the textural features of the input image. It is mainly based on run length. Run length is the frequency of the neighbors having same gray intensity pixels in same direction. GLRLM is a two dimensional matrix. In GLRLM each element p(k,l/0) is the number of elements l with the grey level value k in the direction 0.The"Fig 3(a)"represents a matrix of size 4*4 pixel intensities with 4 gray levels.The"Fig 3(b)".shows GLRLM with 00 direction, GLRLM can also be generated in the directions 45^0,90^0,135^0 as shown in "Fig 3(c)".

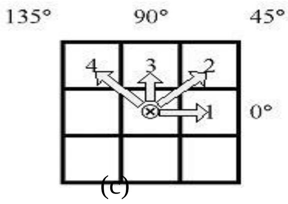
The proposed work extracts 7 texture features from GLRLM namely, short run emphasis(SRE), Long Run Emphasis(LRE), Grey Level Non-uniformity(GLN), Run-Length Non-uniformity(RLN) and Run percentage(RP), low grey-level run emphasis(LGRE) and high grey level run emphasis(HGRE). These features use the pixels grey level in sequence. The purpose of these features is to differentiate the texture having equal value of LRE and SRE. It has differences in the distribution of grey levels.

1	0	0	2	2
3	3	3	0	0
2	2	2	3	3
0	0	0	1	1
0	2	2	3	3

4	1	1	0
1	1	0	0
0	0	4	2
1	0	0	4

(a) (b)

Fig. 2 (a) Image with 4 gray-levelb) GLCM for distance 1 and direction 0^0



(c)

Fig. 2(c) Direction of GLCM generation. From the center to thepixel 1 representing direction = 0^0 with distance d = 1, to the pixel 2direction = 45^0 with distance d = 1, to the pixel 3 direction = $90^0$with distance d = 1, and to the pixel 4 direction = 135^0 with distanced = 1

1	1	3	4
2	3	1	4
2	2	3	2
3	1	4	1

Gray Level	*Run Length(j)*			
	1	2	3	4
1	3	1	0	0
2	2	1	0	0
3	4	0	0	0
4	3	0	0	0

(a) (b)

Fig.3 (a)Image of 4x4 pixels.b) GLRLM Matrix

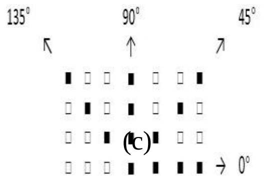


Fig (c)Run Direction

After the calculation of GLRLM features, optimal features are identified using Genetic Algorithm.

C. *Texture features of GLRLM*

TABLE 1: GLRLM FEATURES

S.No	*Features*	*Formulae*
1	Short Run Emphasis(SRE)	$\frac{1}{n}\sum_{i,j}\frac{p(i,j)}{j2}$
2	Long Run Emphasis(LRE)	$\frac{1}{n}\sum_{i,j} j2\,p(i,j)$
3	Gray Level Non-uniformity(GLN)	$\frac{1}{n}\sum_{i}(\sum_{j} p(i,j))2$
4	Run Length Non-uniformity(RLN)	$\frac{1}{n}\sum_{j}(\sum_{i} p(i,j))2$
5	Run percentage(RP)	$\sum_{i,j}\frac{n}{p(i,j)j}$
6	Low Gray Level Run Emphasis(LGRE)	$\frac{1}{n}\sum_{i,j}\frac{p(i,j)}{i2}$
7	High Gray Level Run Emphasis(HGRE)	$\frac{1}{n}\sum_{i,j} i2(i,j)$

Seven texture features can be extracted from the GLRLM. These features uses grey level of pixel in sequence and is intended to distinguish the texture that has the same value of SRE and LRE but have differences in the distribution of gray levels. Once features sets are constructed using GLRLM, then the next section explicates SVM classifier for the classification of extracted features.

D Genetic Algorithm

Like an optimization approach, a set of possible solutions examine and manipulate simultaneously using GA algorithm. For the optimization problems the GA begins with many different solutions which are treated as individuals in a population. These solutions are represented with binary strings by coding process. It is called chromosome. The initial population is constructed randomly. The partitioning specific fitness function is used to evaluate these individuals. The GA produce a new generation of hopefully better solutions by using these individuals. Two of the individuals are selected probabilistically as parents, with selected probability proportional to their fitness in each generation. Crossover is applied on these individuals to produce two new individuals called offsprings, by interchanging parts of their structure. So each offspring acquires a combination of features from both parents. In mutation a least probability incremental modification is made to each member of the population. This indicates that the GA can provide new features. However these features may not be in the population. It makes the total search space reachable in spite the population size. The simplest stochastic selection technique is Roulette wheel parent selection method. In our generation replacement approach, the most inferior member in a population is replaced by new offspring.

When a chromosome represents a selected feature subset, Y, and the evaluation function is clear then the evaluation is straightforward.

The fitness of a chromosome C is defined as

Fitness© = f(YC)

Here f(YC) is the classification error rate.

Here YC is the corresponding feature subset of C

Parameters

The following parameter set is used for GA procedure. These values may be tuned depends data set which may give improved performance.

Control procedure: steady-state

population size=40

pc(crossover probability)=1.0 for steady state(always applied),0.6 for generational

pm (mutation rate)=0.1

q(in rank-based selection)=0.25

Genetic algorithm converges the solution by minimizing the fitness function.

Stopping condition: The difference of two fittest individuals in successive iterations is less than 0.001.Flow Chart is as shown in "Fig.4"

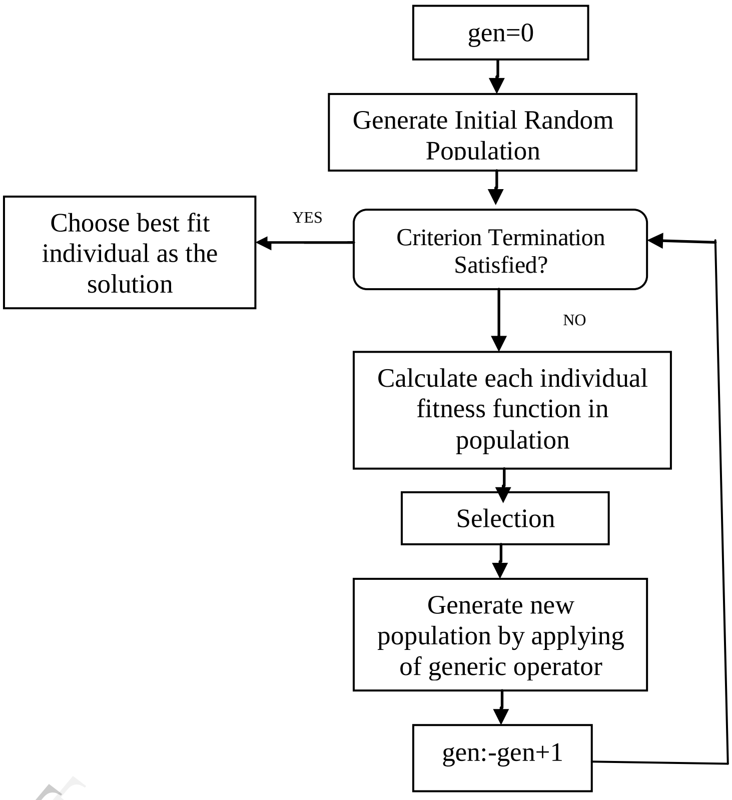

gen=0
Generate Initial Random Population
Criterion Termination Satisfied?
YES
Choose best fit individual as the solution
NO
Calculate each individual fitness function in population
Selection
Generate new population by applying of generic operator
gen:-gen+1

Fig 4.The flowchart of the Genetic Algorithm

III. RESULTS AND DISCUSSIONS

The digital mammograms available in the MIAS database [17] were used for the experiments. The database includes 322 mammograms belonging to normal (Norm) and six abnormal classes—architectural distortion (Arch), asymmetry (Asym), calcification (Calci), circumscribed (Circ) masses, spiculated (Spic) masses and ill-defined (Ill-def) masses. Each mammogram is of size 1,024×1,024 pixels, and annotated for the class, severity, center of abnormality, background tissue character and radius of a circle enclosing the abnormality. The ROIs from abnormal mammograms were extracted. Hence, the abnormal ROIs are of different sizes. But, in the case of normal mammograms, the ROIs of uniform size 200×200 pixels were cropped about the center, which is a new approach that avoids bias in the case of normal mammograms. Out of 322 ROIs, there are 209 normal, 19 architectural distortions, 15 asymmetry cases, 26 calcification regions, 24 circumscribed masses, 19 spiculated masses and 15 ill-defined masses. In this work, all the 327 ROIs were involved to create the feature dataset and 110 ROIs comprising one-third from each mammogram class were selected as queries. The performance analysis of the proposed gray level statistical matrix for texture feature extraction regarding mammogram retrieval problem is presented in this section. Overall Performance offered by various methods is reported.

Table 2: Performance rates of proposed method and competing methods

Method	***Performance(in terms of error rate)***
GLNM	53%
GLAM	63%
GLCM	60%
GLRLM	64.6%
Hybrid GLRLM	64.3%

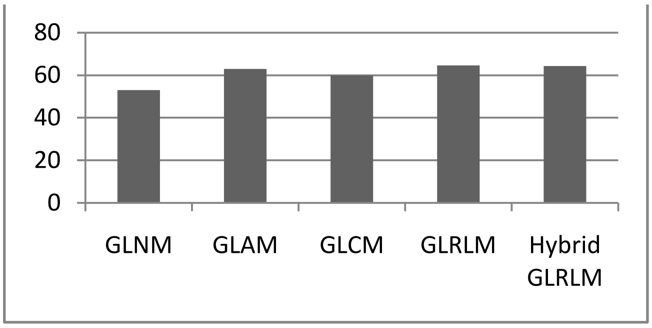


Fig.5Graph for Performance Analysis

IV. CONCLUSION

The paper reports the retrieval performance of the GLRLM and Hybrid GLRLM by applying on MIAS database. The capability of the methods in extracting texture features is demonstrated. These retrieval approaches may help the physicians to effectively search for relevant mammograms during their diagnosis. From the results, comparatively better classification rate is observed in GLRLM and Hybrid GLRLM. Developing more efficient feature estimation method is our future endeavor.

REFERENCES

[1] Majid AS, de Paredes ES, Doherty RD, Sharma N, Salvador X(2003) “Missed breast carcinoma: pitfalls and pearls,” Radio-graphics 23:881–895

[2] Osmar RZ, Antonie M-L, ComanA (2002) “Mammographyclassification by association rulebased classifier,” MDM/KDD2002 International workshop on multimedia data miningwith (ACM SIGKDD 2002, Edmonton, Alberta, Canada, 17–19July 2002), pp 62–69

[3] Xie X, Gong Y, Wan S, Li X (2005) “Computer aided detection of SARS based on radiographs data mining,” In: Proceedings of the2005 IEEE engineering in medicine and biology 27th annualconference Shanghai, China, 1–4 Sept 2005

[4] Shuyan W, Mingquan Z, Guohua G (2005) “Application of fuzzycluster analysis for medical image data mining.,” In: Proceedingsof the IEEE international conference on mechatronics & auto-mation Niagara Falls, Canada, July 2005

[5] Jensen R, Qiang S (2004),“Semantics preserving dimensionalityreduction: rough and fuzzy-rough based approaches,” IEEE TransKnowl Data Eng 16:1457–1471

[6] Walid E, Hakim H (2006),” A new cost sensitive decision treemethod application for mammograms classification,” IJCSNS Int J Comp SciNetwSecur, 6 No. 11

[7] Liu Y, Zhang D, Lu G (2008),” Region based image retrieval withhigh-level semantics using decision tree learning,” Pattern Rec-ognit 41:2554–2570

[8] Polat K, Gunes S (2009), ” A novel hybrid intelligent method basedon C4.5 decision tree classifier and one-against-all approach formulti-class classification problems,” Expert Syst Appl36:1587–1592

[9] Chen C, Lee G (1997), “ Image segmentation using multitiresolu-tion wavelet analysis and expectation maximum (em) algorithmfor mammography,” Int J Imaging SystTechnol 8(5):491–504

[10] Wang T, Karayaiannis N (1998),“Detection of microcalcificationsin digital mammograms using wavelets,” IEEE Trans Med Imaging 17(4):498–509

[11] Christiyanni I et al (2000), “ Fast detection of masses in computeraided mammography,” IEEE Signal Process Mag 54–64

[12] AswinikumarMohanty, ManasRanjanSenapati, Saroj Kumar Lenka(2013), “A novel image mining technique for classification of mammograms using hybrid feature selection,” Neural Comput&Applic (2013) 22:1151-1161.

[13] AswinikumarMohanty, ManasRanjanSenapati, SwapnasiktaBeberta, Saroj Kumar Lenka, “Texture based features for classification of mammograms using decision tree,” Neural Comput&Applic(2013) 23:1011-1017

[14] ManavalanRadhakrishnan and ThangavelKuttiannan, ”Comparative Analysis of Feature Extraction Methods for the Classification of Prostate Cancer from TRUS Medical Images,” IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No 2, January 2012, ISSN (Online): 1694-0814

[15] Kulak E (2002), “ Image analysis of textural Features for Content based retrieval,” Thesis, Sabanci University.

[16] Do MN, Vetterli M (2002), “Wavelet-based texture retrieval using generalized gaussian density and Kullback–Leibler distance,” IEEE Tans Image Proc 11(2):146–158.

[17] Suckling J, Parker J, Dance DR, Astley SM, Hutt I, Boggis CRM, Ricketts I, Stamatakis E, Cerneaz N, Kok SL, Taylor P, Betal D, Savage J (1994), “ Mammographic image analysis society digital mammogram database,” Proceedings of International Workshop on Digital Mammography pp 211-221

[18] Tang X (1998), “Texture information in run-length matrices”, IEEE Trans Image Process 7(11):1602–1609