Diagnose the Stages of Breast Cancer using SVM

Jini.R. Marsilin
PG Student
Department of Computer Science & Engg.
Dr.Sivanthi Aditananr College of Engineering
Tiruchendur-628215

Dr. G. Wiselin Jiji
Professor & Head
Department of Computer Science & Engg.
Dr.Sivanthi Aditananr College of Engineering
Tiruchendur-628215

ABSTRACT
This paper presents a pattern similarity scheme for predicting the real stage of breast cancer. This project allowed the development of content based image retrieval (CBIR) systems, capable of retrieving images based on their similarity with the query image and identifies the correct stages of the breast cancer. The proposed scheme involves low level feature extraction from images like shape and texture features. Shape features used in this scheme are Zernike moments and Radial Chebyshev moments. Texture features of contrast, energy and run length matrix features are also used with the shape features. These extracted features are then classified using SVM. The output of the SVM is considered as patterns. The similarity between two patterns is estimated as a function of the similarity of both their structures and the measure components. The proposed scheme can be effectively applied for image retrieval from large databases and also used to determine the correct stage of breast cancer and get the treatment in appropriate time

Keywords
SVM classifier, Content based image retrieval (CBIR), feature, pattern similarity.

1. INTRODUCTION
Breast cancer is the leading cause of cancer related death in women. Breast cancer is where cancerous (malignant) cells are found in the breast tissue. Getting a mammogram is an effective way to detect breast cancer in its early stages. If we find it in the initial stage, we can cure it.

Globally, 1.3 million women are diagnosed with breast cancer every year. India has one of the highest **deaths per incident ratio**, at almost 50%, compared to 30% in China and 18% in the US. This implies breast cancer is not detected earlier in India, more than any comparable developing or developed country. Common risk factors of breast cancer are: 1. Social: Later age at childbirth, fewer children and shorter duration of breast feeding. 2. Cultural: Fear of self examination, fear of chemotherapy and hair loss, fear of disruption in family responsibilities. 3. Economic: With increasing number of women in the workforce and, the consumption of fatty foods has increased substantially.

This project is intended for reducing the death rate by early detection. The objective of using intelligent system is to predict the real stage of breast cancer. The project will certainly be the most effective and straight forward for the diagnosis of breast cancer stage. Image retrieval is based on similarity measures estimated directly from low level image features.

CBIR refers to the retrieval of images from a database using information derived from the images themselves rather than using text indices. Clustering approach used in [9] offers high retrieval effectiveness with low space overhead. CBIR approach used in [1] utilizes a continuous and probabilistic Gaussian mixture modeling (GMM) along with the Kullback-Leibler (KL) measure. This approach is effectively applied for radiographic images but large image retrieval task's efficiency is not satisfied.

Semantic error–correcting output codes (SECC) based on individual classifiers combination [2], and a framework for medical image retrieval using machine learning and statistical similarity matching techniques with relevance feedback [3], similarity learning approach to content based image retrieval: Application to digital mammography [4] requires prior knowledge about the dataset. These approaches also introduce constraints to the semantics required for image retrieval task.

The proposed scheme use similarity between complex patterns for the image retrieval. This scheme is used to retrieve the images using the pattern similarity from large dataset effectively.

2. STAGES OF BREAST CANCER
The first stage is stage 0, in this stage there are abnormal cells present that might suggest that one is at higher risk for cancer.

- Stage I: where the tumor is less than 2 cm across, and has not spread into the surrounding areas.
- Stage II (A or B): when the cancer is from 2-5 cm across, and has spread into the surrounding areas including the lymph nodes (which must also be removed to prevent the further spread of the cancer)
- Stage III (A, B, or C): It is the advanced stage (i.e.) more than 5 cm across and has spread to the lymph nodes. Cancer is blocking the lymph nodes.
- Stage IV: the cancer has spread the lymph nodes and also spread to other parts of the body like bones, lungs, liver, or brain.

Stages I, IIA, IIB, and IIIA are the "early-stage" breast cancer. This scheme allowed the development of content based image retrieval (CBIR) systems, capable of retrieving images based on their similarity with the query image and identifies the correct stages of the breast cancer.

3. METHODOLOGY

The proposed scheme involves four steps.

- Low-level feature extraction
- Classification – SVM
- Pattern instantiation
- Computation of pattern similarity.

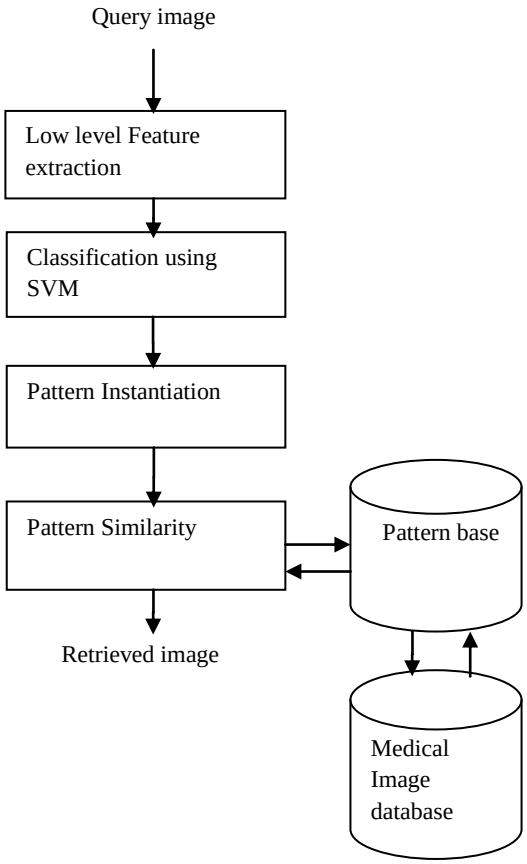

Figure 1: Block diagram of proposed image retrieval system.

From the query image (breast cancer), the low level features are extracted. These extracted features are then classified using SVM Algorithm. Each group is considered as patterns. In pattern instantiation stage, the structural and measure components are represented. Finally, similarity between two patterns is estimated using the distance measures of these components. Using this similarity the most similar images are retrieved with respect to the query image that image is used to identify the real stage of breast cancer.

3.1 Pattern Base (PB):

It keeps the information about extracted patterns from the images. It consists of 3 basic layers. They are,

- pattern type
- pattern
- class

Pattern type defines the description of the pattern structure. Pattern is the instance of the corresponding pattern type. Class is a collection of patterns of the same pattern type. Pattern type PT is defined as a pair PT = {SS, MS} or p={s, m} where SS is the pattern space by describing the structure schema of the pattern type. MS is the quality of the source data representation. A pattern-type PT is called complex if its structure schema SS includes another pattern type, otherwise PT is called simple [6].

3.2. Low-Level Image Feature Extraction:

Features extraction of this similarity scheme describes the process of determining the relevant local content of the image. It includes straight forward function of raw pixel values. Each of the images in the database is scanned with a sliding window of user defined size, sampling image blocks at a given sampling step. The sampling step may allow consecutive blocks to overlap. For each block a set of N features *fi*, *i* = 1. . . *N* is calculated to form a single feature vector *F*. The number of feature vectors depends on the size.

Shape and texture features play an important role in low level image analysis and understanding. Using the moment feature like Zernike moment and Radial Chebyshev Moment (RCM) feature, shape features are extracted. In case of texture, Co-occurrence matrix, run length matrix based schemes are used to extract texture features. Here contrast and energy features use the co-occurrence matrix scheme at the same time run length matrix scheme is used by run length matrix features.

3.2.1. Shape Features:

Shape is the function of position and direction of a simply connected curve within a two dimensional field.

3.2.1.1 Zernike Moments (ZM):

Zernike moments are a set of orthogonal Zernike polynomials defined over the polar coordinate space inside a unit circle $x2 + y2 \leq 1$ [5]. The complex Zernike moments of order n with repetition m of a function f(x, y) are defined as

$$A_{nm} = \frac{n+1}{\pi} \iint_{x^2+y^2 \leq 1} f(x,y).V^*_{nm}(\rho,\theta)dx,dy \text{ ---- (1)}$$

Circular Zernike polynomials in a unit circle are defined as:

$$V_{nm}(\rho,\theta) = V_{nm}(r\cos\theta, r\sin\theta) = R_{nm}(\rho)e^{im\theta} \text{ ----- (2)}$$

ρ is the length of the vector from origin to pixel (x, y) and θ is the angle between vector and the x -axis in the counter-clockwise direction. $x^2+y^2=1$, x =r cos θ y=r sin θ. For a digital image, eq (1) is replaced by,

$$A_{nm} = \frac{n+1}{\pi} \sum_x \sum_y f(x,y).V^*_{nm}(\rho,\theta), x^2+y^2 \leq 1 \text{ ----- (3)}$$

Center of the image is taken as origin & image is mapped inside the unit circle. Pixels that fall outside are neglected. Zernike moments are so powerful because of their insensitivity to image noise, information content and ability to provide faithful image representation. Each Zernike moment acquires a phase shift on rotation. The simple rotational transformation property leads to the conclusion that the magnitudes of the Zernike moments of a rotated image function remain identical to those before

rotation. Thus the magnitude of the Zernike moment can be taken as a rotation invariant feature of the underlying image function. They are robust to noise and minor variations in shape.

3.2.1.2 Radial Chebyshev Moments (RCM):

Chebyshev moments are superior to Zernike moments in terms of image reconstruction capability. Radial Chebyshev Moments also have rotational invariance property. The Radial Chebyshev Moment with order p and repetition q for the image of size- N × N is defined as:

$$S_{pq} = \frac{1}{2\pi\rho(p,m)} \sum_{r=0}^{m-1} \sum_{\theta=0}^{2\pi} t_p(x) e^{-jq\theta} f(x,\theta)$$

------ (4)

where m denotes (*N/ 2) +1*. The recursive relation is defined by,

$$t_0(x) = 1$$

$$t_1(x) = (2x - N + 1)/N$$

$$t_p(x) = \frac{(2p-1)t_1(x)t_{p-1}(x) - (p-1)\left\{1 - \frac{(p-1)^2}{N^2}\right\}t_{p-2}(x)}{p},$$

$$p > 1$$ ----- (5)

(p, N) is given by,

$$\rho(p,N) = \frac{N\left(1-\frac{1}{N^2}\right)\left(1-\frac{2^2}{N^2}\right)\ldots\left(1-\frac{p^2}{N^2}\right)}{2p+1},$$

$$p = 0,1.....N-1$$ ------ (6)

The magnitudes of RCM are also similar to ZM (i.e.) invariant to the rotation [7]

3.2.2 Texture Features:

Texture feature gives the information about the spatial arrangement of color or intensities in an image or selected region of the image.

3.2.2.1 Contrast:

Contrast is the measure of the intensity contrast between a pixel and its neighbor over the whole image.ie. Measures the local variations in the gray –level co-occurrence matrix

$$Contrast = \sum_i \sum_j (i-j)^2 C_d(i,j)$$ ------ (7)

Gray level co-occurrence matrix (GLCM) [8] captures the spatial dependence of gray level values within an image. It characterize the texture of an image by calculating how often pixel with intensity value i occurs in a specific spatial relationship occur in an image to a pixel with intensity value j. C (i, j) indicates how many times value i co-occurs with value j in a particular spatial relationship d.

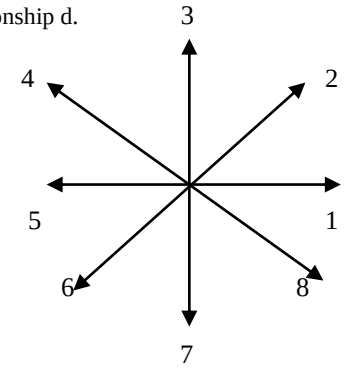

Figure 2: directions for co-occurrence matrix calculus

3.2.2.2 Energy:

It provides the sum of squared elements in the GLCM. It is also known as uniformity or the angular second moment.

$$Energy = \sum_i \sum_j C_d(i,j)^2$$ ------ (8)

3.2.2.3 Run length matrix features:

Run length matrix is defined as the number of runs with pixels of gray level I and run length j. For a given image, a gray level run is a set of consecutive, collinear pixels having the same gray level. Length of the run is the number of pixels in the run.

1. Short Run Emphasis (SRE):

It measures the distribution of short runs.

$$SRE = \frac{1}{n_r} \sum_{i=1}^{M} \sum_{j=1}^{N} \frac{p(i,j)}{j^2} = \frac{1}{n_r} \sum_{j=1}^{N} \frac{p_r(j)}{j^2}$$ ------ (9)

Here M is the number of gray levels, N is maximum run length.

2. Long Run Emphasis (LRE):

It measures the distribution of long runs.

$$LRE = \frac{1}{n_r} \sum_{i=1}^{M} \sum_{j=1}^{N} p(i,j).j^2 = \frac{1}{n_r} \sum_{j=1}^{N} p_r(j).j^2$$

----- (10)

3. Gray-Level Nonuniformity (GLN):

It measures the similarity of gray level values throughout the image.

$$GLN = \frac{1}{n_r} \sum_{i=1}^{M} \left(\sum_{j=1}^{N} p(i,j) \right) = \frac{1}{n_r} \sum_{i=1}^{M} p_g(i)^2$$

------ (11)

4. Run Length Nonuniformity (RLN):

It measures the similarity of length of runs throughout the image.

$$RLN = \frac{1}{n_r}\sum_{j=1}^{N}\left(\sum_{i=1}^{M} p(i,j)\right)^2 = \frac{1}{n_r}\sum_{j=1}^{N} p_r(i)^2 \quad \text{-------- (12)}$$

5. Run Percentage (RP):

It measures the homogeneity and the distribution of runs of an image in a single direction.

$$RP = \frac{n_r}{n_p} \quad \text{-------- (13)}$$

6. Low Gray-Level Run Emphasis (LGRE):

It measures the distribution of low gray level values.

$$LGRE = \frac{1}{n_r}\sum_{i=1}^{M}\sum_{j=1}^{N}\frac{p(i,j)}{i^2} = \frac{1}{n_r}\sum_{i=1}^{M}\frac{p_g(j)}{i^2} \quad \text{-------- (14)}$$

7. High Gray-Level Run Emphasis (HGRE):

It measures the distribution of high gray level values.

$$HGRE = \frac{1}{n_r}\sum_{i=1}^{M}\sum_{j=1}^{N} P(i,j).i^2 = \frac{1}{n_r}\sum_{i=1}^{M} p_g(j).i^2 \quad \text{------- (15)}$$

8. Short Run Low Gray-Level Emphasis (SRLGE):

It measures the joint distribution of short runs and low gray level values.

$$SRLGE = \frac{1}{n_r}\sum_{i=1}^{M}\sum_{j=1}^{N}\frac{P(i,j)}{i^2.j^2} \quad \text{------- (16)}$$

9. Short Run High Gray-Level Emphasis (SRHGE):

It measures the joint distribution of short runs and high gray level values.

$$SRHGE = \frac{1}{n_r}\sum_{i=1}^{M}\sum_{j=1}^{N}\frac{P(i,j).i^2}{j^2} \quad \text{------- (17)}$$

10. Long Run Low Gray-Level Emphasis (LRLGE):

It measures the joint distribution of long runs and low gray level values.

$$LRLGE = \frac{1}{n_r}\sum_{i=1}^{M}\sum_{j=1}^{N}\frac{P(i,j).j^2}{i^2} \quad \text{------- (18)}$$

11. Long Run High Gray-Level Emphasis (LRHGE):

It measures the joint distribution of long runs and high gray level values.

$$LRHGE = \frac{1}{n_r}\sum_{i=1}^{M}\sum_{j=1}^{N} P(i,j).i^2.j^2 \quad \text{------ (19)}$$

3.3. Classification- SUPPORT VECTOR MACHINES:

Support vector machines (SVM) are binary classifiers that estimate the optimum separating hyper plane that maximizes the margin between two classes. The margin can be defined as the distance of the closest point, in each class, to the separating hyper plane. It classifies both linear and nonlinear data. Support vectors (SVs) contain highlighted pixels that help to create the margins or boundaries in an image.

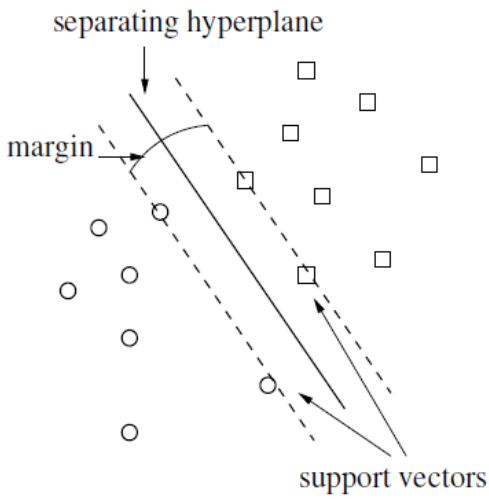

Figure 3: linear SVM

Training set is transformed to higher dimension by Nonlinear mapping. Within this, it searches for the decision boundary linearly separating the data of one class from another. Each training tuple in the data set is associated with class label. This label can take one of two values, either +1 or -1. If the data is linearly separable means a straight line is drawn to separate all the tuples of class +1 from all the tuples of class -1. Larger margin hyper plane is more accurate than the hyper plane with smaller margin. In the training phase, the identified support vectors are used to classify each pixel of an image according to the class predicted by the support vector machine classifier.

3.4. Pattern Instantiation:

The results from the SVM algorithm are considered as patterns extracted from the image database.

Each class of image comprising of *M* simple patterns P_i , $i = 1. . . M$, and with respect to the output of the SVM algorithm, a Specimen $_i$ is represented for each pattern P_i of a medical image

Specimen $_i$ =

$$\begin{pmatrix} SS:(D:[\mu:[\mathrm{Real}],\sigma:[\mathrm{Real}]]_1^N) \\ MS:(pp:[\mathrm{Real}],SV:[\mathrm{Real}] \end{pmatrix} \quad \text{---- (20)}$$

Where Structure schema SS is represented by the pair (μ, σ) of the distribution *Dj* in pattern P_i. Measure schema MS is represented by two values, the *prior probability* (*pp*) and the *scatter value* (SV) of P_i. Prior probability *pp* is defined as the fraction of the feature vectors of the image that belong to pattern P_i. In this case, it provides an indication of the size of the specimen. SV is a measure of the cohesiveness of the data items in a group with respect to the centroid of the group, and it is a commonly used intrinsic measure of the quality of a cluster [8]. If the SV is low, it indicates good scatter quality.

3.5. Pattern Similarity:

Pattern similarity is estimated by using the distance over the structural and the measure components of two simple patterns *P*1 and *P*2. Complex patterns are decomposed into a number of simple patterns. When comparing two medical images, MI1 and MI2, there is a need to associate component patterns of MI1 to component patterns of MI2.

First, find the distance of measure components using absolute difference of the scatter values, each one weighted by the corresponding prior probability of the patterns, normalized by the sum of the two scatter values.

$$dis_{meas}(P_1, P_2) = \frac{|P_1.pp.P_1.SV - P_2.pp.P_2.SV|}{P_1.SV + P_2.SV} \quad \text{------ (21)}$$

For finding structural similarity between *P*1 and *P*2, first find the standardized difference d between two distributions by Cohen's distance metric.

$$d(D1,D2) = \begin{cases} \dfrac{|D_1.\mu - D_2.\mu|}{\sqrt{\dfrac{D_1.\sigma^2 - D_2.\sigma^2}{2}}}, & if\ D_1.\sigma \neq 0\ \ or\ \ D_2.\sigma \neq 0 \\ |D_1.\mu - D_2.\mu|, & otherwise \end{cases} \quad \text{------ (22)}$$

If d=0, then distributions are identical. Low d value refers quite similar distributions and high d value refers quite dissimilar distributions.

Structural distance between two sets of distributions should be the result of aggregate function.

$$dis_{struct}(P_1, P_2) = g_{aggr}\left(\frac{d(D_j^1, D_j^2)}{\delta}\right) \forall j = 1.2..N \quad \text{------ (23)}$$

Distance between two patterns dis (p1, p2)

$$dis(P_1, P_2) = dis_{struct}(P_1, P2) + (1 - dis_{struct}(P_1, P_2)).dis_{meas}(P_1, P_2)^2 \quad \text{------ (24)}$$

To compare two medical images MI1, MI2 (complex patterns), distance between the different patterns of each image is used.

$$\text{dis}(\text{MI}_1, \text{MI}_2) = \frac{1}{M.K}.\sum_{i=1}^{M}\sum_{j=1}^{K} dis(P_i^{MI_1}, P_j^{MI_2}) \quad \text{------ (25)}$$

M &K are the numbers of constituent simple patterns of each image. The final outcome is the average of all possible matching. The retrieved image is similar to the query image that gives the real stage of the breast cancer.

4. RESULTS

The experiments are performed with the selected mammogram images taken randomly from patients of different ages and pathologies during medical routine. The sampling parameters tested before each CBIR experiment includes sliding windows of 32 × 32, 64 × 64 and 128 × 128 pixels. Increasing the overlap provides better localization of the patterns but produces many more sampled blocks, affecting the efficiency of both the feature extraction and the pattern instantiation tasks. Thus, a 50% overlap, i.e. a 32 pixel step, was used as a compromise between localization and efficiency.

In order to quantitatively assess the effectiveness of the proposed pattern similarity scheme, evaluate its capability to retrieve images by adopting the popular *recall* and *precision* measures. Precision is the faction of retrieved images that are relevant. Recall is the fraction of images that are relevant to the query that are successfully retrieved. Area under the interpolated precision-recall curve (AUC) expected for the proposed approach reached 78% whereas AUC estimated from [1] is approximately 66%.

5. CONCLUSION

This approach is applied for efficient content based image retrieval. Query image is compared with the medical image database and the relevant image is retrieved using pattern similarity scheme. This scheme has the capability of coping with large image retrieval tasks. The real stage of the breast cancer is retrieved from this approach.

6. REFERENCES

[1] H. Greenspan and A. T. Pinhas, "Medical image categorization and retrieval for PACS using the GMM-KL framework," IEEE Trans. Inf. Techol. Biomed. vol. 11, no. 2, pp. 190–202, Mar. 2007.

[2] J. Yaoa, I, S. Antani, R. Longb, G. Thoma, and Z. Zhanga, "Automatic medical image annotation and retrieval using SECC," in Proc. 19th Int.Symp. Comput.-Based Med. Syst. (CBMS), Salt Lake City, UT, Jun. 2006.

[3] M. M. Rahman, P. Bhattacharya, and B. C. Desai, "A framework for medical image retrieval using machine learning and statistical similarity matching techniques with relevance feedback," IEEE Trans. Inf. Technol Biomed., vol. 11, no. 1, pp. 58–67, Jan. 2007.

[4] I. El-Naqa, Y. Yang, N. P. Galatsanos, R. M. Nishikawa, and M. N. Wernick, "A similarity learning approach to content-based image retrieval: Application to digital mammography," IEEE Trans.

Med. Imag., vol. 23, no. 10, pp. 1233–1244, Oct. 2004.

[5] M. Emre Celebi and Y. Alp Aslandogan "A Comparative Study of Three Moment-Based Shape Descriptors"

[6] Ilaria Bartolini, Paolo Ciaccia, Irene Ntoutsi, Marco Patella, and Yannis Theodoridis "A Unified and Flexible Framework for Comparing Simple and Complex Patterns"

[7] Mukundan R. (2004) "A New Class of Rotational Invariants Using Discrete Orthogonal Moments" Proceedings of the 6th IASTED Conference on Signal and Image Processing, pp. 80-84

[8] Dan Popescu, Radu Dobrescu, and Maximilian Nicolae, "Texture Classification and Defect Detection by Statistical Features".

[9] Renato 0. Stehling Mario A. Nascimento Alexandre X. Falcgo "An Adaptive and Efficient Clustering-Based Approach for Content-Based Image Retrieval in Image Databases"