Received: 23 March 2023 | Revised: 20 July 2023 | Accepted 27 October 2023

Research Article
Journal of Research and Review in Science
94-107, Volume 10, December 2023
DOI: 10.36108/jrrslasu/4202.11.0110

ORIGINAL RESEARCH

JOURNAL OF RESEARCH AND REVIEW IN SCIENCE
LAGOS STATE UNIVERSITY
FOR TRUTH AND SERVICE

Investigating the Effect of Some Selected Distance Measures on the Performance of Query-By-Image-Content Over Mammogram Images

Ronke Babatunde[1], Ayodele Oloyede[2], Temitayo Fagbola[3], Nwaocha Vivian[4], Tola Ajagbe[5], Rilwan Shanu[6].

[1]*Department of Computer Science Kwara State University, Malete, Kwara State, Nigeria*

[2,5,6]*Department of Computer Science, Lagos State University, Ojo, Lagos, Nigeria*

[3]*Department of Computer Science, Federal University, Oye-Ekiti, Nigeria*

[4]*Department of Computer Sciences, Faculty of Sciences, National Open University of Nigeria, (NOUN) Nigeria*

Correspondence
Ayodele Oloyede, Department of Computer Science, Faculty of Science, Lagos State University, Nigeria.
Email: ayodele.oloyede@lasu.edu.ng

Funding information
Grant sponsor and grant number: NUC Research Grant (nuc.edu.ng)
(NUC1311/1)

Abstract:
Introduction: Generating signatures of images and comparing them with those stored in a database for the purpose of retrieval of similar content is termed Query by image content. This art is helpful for detection of various diseases such as breast cancer, brain tumor and spine disorder. The image data are obtained through Computerized Tomography scan, Magnetic Resonance Imaging and mammogram, and are misunderstood in routine screening of mammograms despite the recent scientific improvements.
Aim: This study aims to determine the most suitable distance measure for retrieval of similar images by experimenting state-of-the-art distance measures on mammographic images.
Materials and Methods: The procedure was benchmarked on mini mammographic image analysis society (mini-MIAS) and breast cancer digital repository datasets. The experimental process includes thresholding and extraction of Region of Interest using gray level co-occurrence matrix. The extracted features were tested on Euclidean distance, Minkowski distance, Hamming distance, Mahalanobis distance, Cosine Similarity and Manhattan distance measures. The performance of the system on the distance measure was evaluated on the datasets to determine the distance metric that best identify abnormality in the samples.
Results: The empirical results reveal that Mahalanobis distance measure outperforms the others in terms of retrieval time (1.26s and 1.14s) and minimal error (0.004 and 0.002) respectively for both the mini-MIAS dataset and the BCDR dataset, based on the similarity of images retrieved compared to queried images.
Conclusion: Researchers can take appropriate decision on the choice of distance measure based on these findings

To Keywords: Breast-cancer, Brain-tumor, Spine-disorder, Tomography-scan, Magnetic-Resonance-Imaging.

All co-authors agreed to have their names listed as authors.

1. INTRODUCTION

Query by image content (QBIC) also known as content-based image retrieval has several applications particularly in the medical field. The large quantity of medical images produced in hospitals and health centers calls for a requirement of novel devices to recover the visual information. In recent times, the role of QBIC in the medical field cannot be underestimated as it is a system that assist in the efficient delivery of task [1]. One objective is to present a radiologist with a set of images from earlier cases which are similar to the one estimated, together with the recognized pathology of earlier ones to help in complex situations. As researched by [2],

QBIC can be applied to radiology and outcomes from similar images of cases earlier treated can assist the radiologist to improve interpretation of rare abnormalities and determining diagnosis. Manually interpreting medical images in large volumes consumes a significant amount of time, is a monotonous process, and is vulnerable to mistakes and biases due to the nature of human judgments. QBIC in medical imaging suggests the utilization of a Region of Interest (ROI), texture as well as visual features for effectual retrieval of content. Through textural analysis, it is possible to discover the texture signature of a medical image relating to diagnostic issues.

The efficiency of textural analysis is based on the techniques used to extract significant characteristics of the image. There are various types of textural features, including gray level co-occurrence matrices, local binary pattern and Tamura's textual features [3]. The features of human breast tissues can best be extracted and analyzed using texture analysis. Due to large number of sample images generated in the hospitals, labeling the images and categorizing the abnormality by human involvement is a time consuming and complex task.

Hence, QBIC system can be used to ease the retrieval task from large databases in order to detect abnormalities. Exact classification of images is a vital part of query-based image retrieval from a large database. Recovery of these images depends on similarity matching actions between features of query image with the database images features. The initial step in the technique is preprocessing as noise or misalignment may have occurred in the course of capturing the images. Next is feature extraction in which content of image is denoted by means of set of features to avoid large input to the processing device and eliminate the content feature that does not enclose the helpful information. Feature extraction process removes the unnecessary features and provides the required image. This helps to minimize unnecessary features [4]. Lastly is the similarity measure stage, where a test image is queried, the image similar to the query image is identified and retrieved in the image retrieval process.

The implementation of an image retrieval system first extracts a combination of shape, texture and edge features from an image, depending on the preprocessing and feature extraction technique used. Then weights are generally assigned to each piece of information extracted from the images and an overall similarity is computed. Images can then be ranked based on this similarity computation. QBIC recovers the mammogram images based on the breast tissue density of the query mammogram. It supports the radiologist for effective diagnosis and decision making.

Therefore, QBIC system does not aim to replace the physician by predicting the disease of a particular case but to assist him/her in diagnosis. By checking with the output of a QBIC system, the doctor increases the assurance in decision or even takes other possibilities. Nevertheless, such system poses two difficulties. First is how to identify the ROI as doubtful regions with very weak environment and second is how to remove features which differentiate the doubtful regions [5].

QBIC is a way to index or find a similarity between images in a database. The matching process between image search model and accumulated image content measures are complex and require sophisticated data management support. Recovery by image content has been a research area of huge interest in the last decade. Several techniques have been proposed to the problem of finding or indexing images based on their contents and each method used has strong and weak points. Over several years various techniques for image retrieval have been used and implemented but QBICstill suffers from challenges such as semantic gap and varied interpretation of visual data by different users [6].

However, the general approach to QBIC represent each image in the database by a vector of feature values [7]. The distance metrics gives the similarity index between two vectors when compared. A number of Machine Learning Algorithms - Supervised or Unsupervised, use Distance Metrics to know the input data pattern in order to make any Data Based decision. A good distance metric helps in improving the performance of Classification, Clustering and Information Retrieval process significantly.

Ideally, distance metrics use a distance function that tells us the mathematically driven distance between various elements throughout the data set. The closer the distance, the more similar they are and vice-versa. There are several measures of distance that can be used, and it is important to be aware of them while considering the best solution for a given situation to avoid errors and interpretation issues. The commonly used Distance metrics in Machine learning according to [8] are: Euclidean Distance, Minkowski Distance, Manhattan Distance, Mahalanobis Distance, Hamming distance and cosine similarity. Therefore, the choice of an appropriate distance metrics for a problem domain will result in an efficient classification system.

In this paper, a comparison of the aforementioned standard distance metrics used in machine learning was carried out on a query by image content system. The rest of this paper is organized as follows. Section 2 discusses the allied research in the field of QBIC. Section 3 explains the procedure and experimental analysis, while section 4 discusses the result of the experiment. Section 5 concludes the work.

2. RELATED WORKS

The query by image content has proved to be an efficient image retrieval system for retrieving the database images that exhibit similarity to the query image presented by the user. [9] developed an interactive content-based image retrieval using ripplet transform and fuzzy relevance feedback. The CBIR system was derived from a new Multiscale Geometric Analysis (MGA) tool named as Ripplet Transform Type-I (RT). To enhance the retrieval performance, a Fuzzy Relevance Feedback Mechanism (FRFM) was designed. Fuzzy entropy-based feature evaluation mechanism was employed for automatic calculation of revised feature’s significance and correspondence distance at the end of all iteration, yielding an improved performance.

[10] designed a framework for medical image retrieval using merging-based classification with dependency probability-based relevance feedback. A novel fitness function was designed to place related images in the database together in the feature space. By means of the merging-based classification, the m-nearest classes to the query image were chosen as a filtered search space. To enhance the retrieval efficiency, a dependency probability-based relevance feedback approach was combined with the QBIC framework. The result obtained shows the efficiency of the method.

[11] proposed a color directional local quinary patterns for content-based indexing and retrieval. A developmental approach was designed to remove color-texture features for image retrieval application called Color Directional Local Quinary Pattern (CDLQP). The method removes the individual R, G and B channel wise directional edge information between reference pixel and its neighborhood by calculating its grey-level difference depending on quinary value (−2, −1, 0, 1, 2). The result shows comparable performance with existing model.

[12] developed a computer aided detection algorithm for digital mammogram images. Preprocessing, enhancement, feature extraction, segmentation and classification were carried out. Segmentation was done using K means clustering algorithm and then feature extracted by gray level co-occurrence matrix. Classification was performed by SVM, yielding successful simulation result. The system however was limited by slow retrieval time.

[13] carried out a review of common approaches to QBIC system which includes color moment, Gabor filters and Fourier transforms. The work identified that the use of low-level spatial features (visual features)

including shape, color, as well as texture features may not be sufficient enough for building robust QBIC system especially when dealing with large databases.

[14] developed a QBIC system using a two-step strategy. It involved designing the first step to extract the low level or pixel level features of the image by using color, texture and shape descriptor. It uses an efficient fused feature extraction method based on color and edge directivity descriptor (CEDD) for extracting the color as well as texture features. A two- level discrete wavelet transform (2D-DWT) was used for extracting the shape feature of the image. Then, in the second step the SVM classifier was used to classify the images into different classes and to handle irrelevant examples. For retrieving the similar images based on query image the Euclidean distance similarity measurement was used. The fused and classifier based proposed scheme was applied on different image databases and it yielded better results over various existing methods and individual approaches.

[15] developed a QBIC system employing metaheuristic algorithm (MA) for the retrieval of images from databases. The proposed QBIC extracts features such as color signature, shape and texture from the image once a query image is entered. Meanwhile, the MA based similarity measure is used to efficiently retrieve images relevant to the query image. The MA used in the research was genetic algorithm (GA) and iterative local search (ILS). The experiments were conducted according to the Corel image database. The work revealed that MA algorithm possesses strong capacity in distinguishing color, shape and texture features. The addition of the ILS algorithm with the GA raised the quality of solution (weight) via the increase in the fitness number, thereby improving the exploitation process in the course of searching. The experimental result showed that the system outperformed state of the art QBIC systems.

[15] conducted a survey of QBIC techniques and challenges faced by QBIC system. The work asserts that despite the vast amount of research carried out so far in the domain, QBIC still faces challenges such as subjectivity of human perception of visual content. There also exists gap between information extracted automatically from visual data and interpretation by the user known as semantic gap. Furthermore, the current QBIC systems still lack accuracy of relevant images due to improper selection of feature extraction methods and similarity measurement technique.

[16] reviewed the state-of-the-art methodology of content-based image retrieval. The various procedure for extracting the image visual substance which includes colour, shape and texture features was examined. The retrieval of similar images using distance measures, Relevance Feedback (RF) and indexing methods were also investigated. The review concluded by stating that the current contemporary procedures used in the design of QBIC systems needs to be improved upon.

Considering the fact that a real time retrieval of similar images is pertinent to the domain in which QBIC is been applied and the adoption of a similarity metrics which could achieve such is germane and calls for continual research.

The literature amasses several techniques used by researchers for QBIC system. However, the use of distance metrics has not been sufficiently evaluated and tested. Thus, a comparison of the performance of the commonly used distance metrics on mammogram images is experimented on two public datasets

3. MATERIAL AND METHODS

Medical digital images contain an immense proportion of data, and their interpretation requires efficient image processing and analysis. In this research, the mammogram data acquired was preprocessed and salient features, which are relevant for further analysis of the QBIC system performance was carried out as outlined in this section.

The high-level diagram of the QBIC system is shown in Figure 1. The grey level co-occurrence matrix of the binarized images were obtained and stored as feature vectors. These feature vectors were compared to each queried image using each of Manhattan, Euclidean, Mahalanobis, Cosine Similarity, Hamming distance and Minkowski distance measure. This was done to compare the performance of the distance metrics on retrieval of similar images based on the image content.

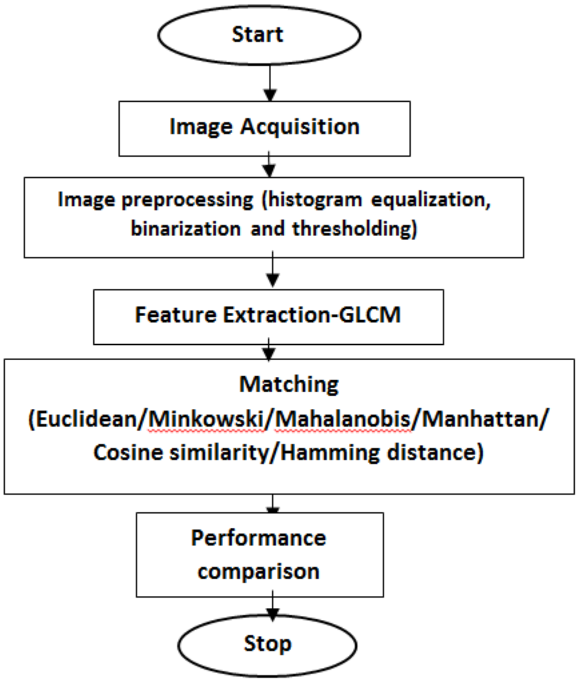

Start
Image Acquisition
Image preprocessing (histogram equalization, binarization and thresholding)
Feature Extraction-GLCM
Matching (Euclidean/Minkowski/Mahalanobis/Manhattan/ Cosine similarity/Hamming distance)
Performance comparison
Stop

Fig. 1: High level diagram of the QBIC system.

3.1 Image data acquisition

In this research, Mammogram data was obtained from the Mini-MIAS database of mammograms at https://www.kaggle.com/kmader/mias-mammography and the breast cancer digital repository (BCDR) at https://bcdr.eu›download. *The MIAS* dataset consist of a total of three hundred and thirty-four (334) sample images which were grouped as having 6 attributes namely; Background Tissue, Class of Abnormality, Severity of Abnormality, x-image coordinate, y-image coordinate and Radius of circle enclosing abnormality. The mammography status conclusion was categorized into benign and malignant. The BCDR database contains 1493 images both from mediolateral oblique (MLO) views and craniocaudal (CC) views, also categorized into benign and malignant.

3.2 Image Preprocessing

In this research, the image pre-processing stage consists of image histogram equalization, binarization and region isolation.

3.2.1 <u>Histogram equalization</u>

Due to the low contrast of medical images which is basically as a result of age of capturing equipment and poor illumination conditions, there exist the need for contrast enhancement before the images are been used for automated systems [17]. Contrast limited adaptive histogram equalization (CLAHE) is a widely used spatial domain enhancement method and is adopted in this research to equalize the non-uniform image due to external conditions to a uniform variation. The “adapthisteq” function in MATLAB was used for the histogram equalization in this research.

3.2.2 Binarization

Image binarization is the process of separation of pixel values into two groups: white as background and black as foreground, where thresholding plays a major role in binarization of images. Image Binarization converts an image in 0 to 255 grey levels to a black and white image. For the equalized image, the pixels are represented in a 0 to 255 grey level intensity to reveal the affected region in the image representation. An arbitrary threshold value chosen will enable the classification of all pixels with values above this threshold as white, and all other pixels as black. The Otsu Thresholding function in MATLAB was used to automatically find an optimal threshold based on the observed distribution of pixel values for the equalized image. The output of the thresholding operation is a binary image [17].

3.2.3 Region Isolation

The *roifilt2* function in MATLAB was applied on the binarized image and subsequently, the filling operator was used to fill the close contours and further, centroids are computed to localize the regions. The ROI, specifies the image to be filtered, the mask that defines the ROI, and the filter that is to be used.

3.3. Feature Extraction and Matching

Feature extraction is the procedure of data transformation and reduction to find a subset of salient variables from the image. In this work, textural features based on the grey level co-occurence matrix (GLCM) are extracted from each mammogram image. The procedure of extraction of GLCM feature used in [17] was adopted. The Co-occurrence matrix is a matrix of two dimensions of joint probabilities between two pairs of pixels which are calculated for four directions: 0º, 90º, 45º and 135º degrees, indicating vertical, horizontal, right and left diagonals respectively. The GLCM features including Standard Deviation, Correlation, Homogeneity, Entropy, Average, Contrast Dissimilarity and Energy, were extracted for every block of the image patch computed in each of the four angles in MATLAB simulator.

These GLCM features extracted from the dataset of images were used for indexing. For each image a 2-D histogram of its hue saturated value (HSV) values was computed. At the end of the indexing stage, all 2-D HSV histograms are stored in the same **.**mat file to create a knowledge base. The retrieval of similar images follows based on the distance metrics as earlier mention, the Manhattan, Minkowski, Mahalanobis, Hamming, Cosine similarity and Euclidean distance.

3.3.1 Manhattan Distance

The Manhattan Distance function calculates the distance traveled if a grid-like path is taken, to get from one data point to another. This is the absolute distance between two points x and y. Therefore, the Manhattan distance is computed as the sum of the variations between the two feature points. This is given in equation 1 as: $d(x,y) = \sum_{i=1}^{n} |x_i - y_i|$ (1)

3.3.2 Minkowski distance

Minkowski Distance is a metric inside a vector space that is normed. This distance metric is used for nearness variable distance to find the similarity of distances between vectors given two or more vectors. Essentially, distance metrics from the Minkowski equation are applied to machine learning to determine the similarity of size. The Minkowski distance between two points x and y is defined as:

$$D = [\sum_{i=1}^{n} |x_i - y_i|^{x}]^{1/x} \qquad (2)$$

3.3.3 Mahalanobis Distance

The Mahalanobis distance is a very useful way of determining the "similarity" of a set of values from an unknown sample to a set of values measured from a collection of "known" samples. The Mahalanobis distance is measured in terms of standard deviations from the mean of the training samples, therefore, the reported matching values is expected to give a statistical measure of how well the spectrum of the unknown sample matches (or does not match) the original training spectra. Thus, Mahalanobis distance can be seen as the generalization of Euclidean distance, and can be computed for each cluster if the covariances of the cluster are known i.e. The Mahalanobis distance between two points x and y is defined as:

$$D_i^{Mahalanobis} = \sqrt{(x - yi)^T \sum_i^{-1} (x - yi)} \qquad (3)$$

where $\sum_i^{-1}$ represents the inverse of the covariance matrix of class i given by

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn} \end{pmatrix}$$

3.3.4 Euclidean distance

Euclidean Distance is the most popular distance metric. Most machine learning algorithms use Euclidean Distance to measure the similarity between observations when dealing with 2 dimensions.

The Euclidean Distance between two points x and y is defined as:

$$D = \sqrt{((x_1 - y_1)^2} + (x_2 - y_2)^2) \qquad (4)$$

3.3.5 Hamming Distance

Hamming distance calculates the distance between two binary vectors, also referred to as binary strings. Hamming distance is a metric for comparing two binary data strings. While comparing two binary strings of equal length, hamming distance is the number of bit positions in which the two bits are different. The distance between two points can be calculated as the sum or the average number of bit differences between the two-bit strings. $D_H = \sum_{i=1}^{k} |x_i - y_i|$ (5)

$$x = y => D = 0$$
$$x \neq y => D = 0$$

3.3.6 Cosine similarity

Cosine similarity is the cosine of the angle between two n-dimensional vectors in an n-dimensional space. It is the dot product of the two vectors divided by the product of the two vector's lengths (magnitudes). It measures the cosine angle between the two vectors.
Cosine similarity ranges from 0 to 1, where 1 means the two vectors are perfectly similar.
If the angle between two vectors increases then they are less similar. The similarity between two vectors A and B is given by $Similarity\ (A, B) = \frac{A*B}{||A||*||B||} = \frac{\sum_{i=1}^{n} Ai*Bi}{\sqrt{\sum_{i=1}^{n} A_i^2} * \sqrt{\sum_{i=1}^{n} B_i^2}}$ (6)

For each 3D *hist* bin, the six different distances (D) were computed between the *hist* of the query image and the i-th database image in the indexed database. The similarities are obtained in a vector and then sorted to prompt the user with the images that have the smallest value as the matching image.

4. EXPERIMENTAL ANALYSIS

The mammogram images were loaded into the developed software GUI interface in MATLAB by using the function "*imread*". This function reads the image from the specified path and displays it into specified axes on the software GUI interface. The syntax for selecting an image is *imread(filename, format)* which reads a grayscale or color image from the file specified by the string filename.

4.1 Result of Preprocessing

Preprocessing helps to balance the intensity difference between the image and background. This result in a reliable representation of the breast tissue structures to reveal the texture and features of the data. Figure 2 shows the outcome of preprocessing activity.

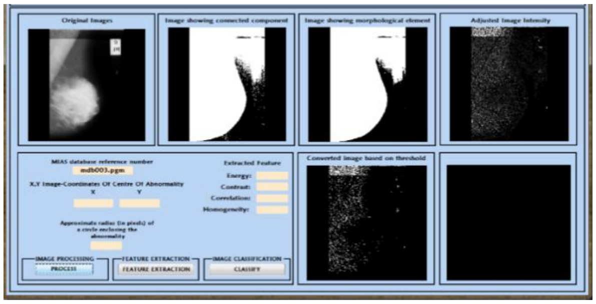

MIAS database reference number: mdb003.pgm
Original Images
Image showing connected component
Image showing morphological element
Adjusted Image Intensity
Extracted Feature
Energy:
Contrast:
Correlation:
Homogeneity:
X,Y Image-Coordinates Of Centre Of Abnormality
X
Y
Approximate radius (in pixels) of a circle enclosing the abnormality
Converted image based on threshold
IMAGE PROCESSING
PROCESS
FEATURE EXTRACTION
FEATURE EXTRACTION
IMAGE CLASSIFICATION
CLASSIFY

Fig. 2: Image Processing

The grayscale mammogram image are further binarized, resulting into image consisting of only two colors namely black and white as shown in figure 3a. The MATLAB function,*BW=bwlabel(image, 8)*was used. This label the connected component in the 2-D image as shown in figure 3b after which the set of properties specified by the properties for each connected component in the binary image, BW are measured and stored in a structured array with the syntax *j=regionprops(BW)*

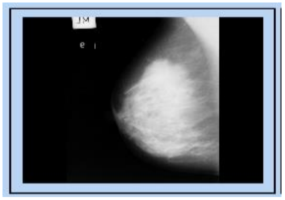


Fig. 3a: Original Mammogram image.

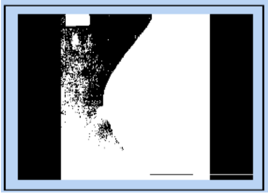


Fig. 3b: Connected component label

The function '*bwareaopen(image, number of pixel)*' was used to remove from the binary image all connected components that have fewer than the number of pixels, producing another binary image. The function '*strel(shape, parameters)*' was used to create a structuring element of type specified by shape (disk, rectangle, diamond, pair, line, octagon etc.). The dark patches or holes present in the image are filled using the function "imfill( )".

A hole is a set of background pixels that cannot be reached by filling in the background from the edge of the image. Finally "*imerode ( )*" function was used to erode the image by structuring element. The boundary of an image is obtained by subtracting the main image and the eroded image. The syntax used here is *IM2=imerode(IM,SE)* which erodes the grayscale, binary, or packed binary image IM, returning the eroded image IM2 The function '*imopen(image, structuring element)*' performs morphological opening on the grayscale or with the single structuring element. The function '*imadjust(image)*',maps the intensity values in grayscale image to new value such that 1% of data is saturated at low and high intensity of the image. This in turn increase the contrast of the output image as shown in figure 4.

.

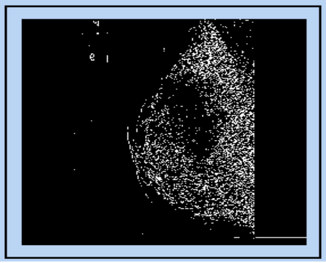


Fig. 4: Adjusted image intensity

The function ‘*graythresh(adjusted image)*’, compute the global threshold level that can be used to convert an intensity image to a binary image. The level is a normalized intensity value that lies in the range of [0, 1]. The ‘*graythresh*’ function uses the Otsu’s method which choose the threshold to minimize the intra class variance of the black and white pixels. The function ‘*im2bw (converted image, threshold level)*’ convert the grayscale image to a binary image. The output image replaces all pixels in the input image with luminance greater than the threshold level with the value 1 (white) and replace all other pixels with the value 0 (black) as shown in figure 5.

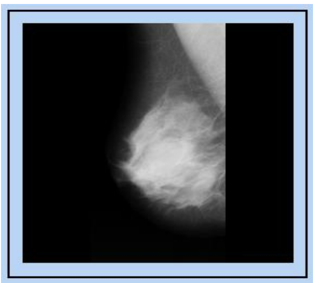


Fig. 5: Binary image base on threshold level

4.2 Obtaining the ROI

The next stage of the mass detection is to separate the suspicious regions that may contain masses from the background parenchyma, i.e., to partition the mammogram into several non-overlapping regions, then extract regions of interests (ROIs), and locate the suspicious mass candidates from ROIs. The suspicious area is an area that is brighter than its surroundings, has almost uniform density, has a regular or irregular shape with varying size, and has fuzzy boundaries. The images are segmented using sub-image GLCM histogram with the function ‘*segmentationGLCM(inputImage, nClusters, isGrayScale)*’. The function implements a rather primitive scheme for image segmentation. Each sub-image is processed using "*blockproc*" to acquire a feature vector for the central sub-image pixel. The feature vector is generated via wrapped "*graycomatrix*" GLCM.As a result, each pixel is clustered resulting in image segmentation as shown in figure 6a and 6b respectively. The ROI helps to reduce the features considered for matching.

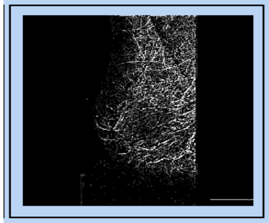


Fig. 6a: blockproc GLCM sub-image

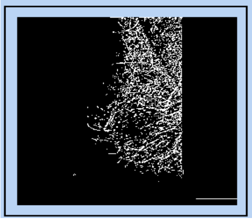
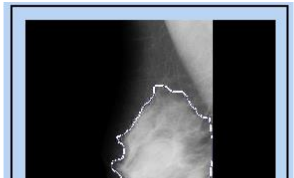
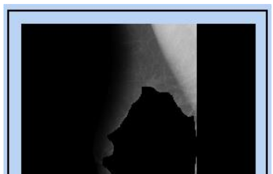


Fig. 6b: graycomatrix sub-image

The graycomatrix sub-image value are then passed to ROI. The ROI locate the difference in GLCM value of the texture image and locate the region where the value are on the entire image to create a mask on the located area with the use of the function ‘cr = *createmask(position, double(inputimage))*’, then the image masked area are carved out by finding the differences between the original image and the mask area with the function ‘*mi = input image - cr’* as shown in figure 7a, 7b and 7c respectively.





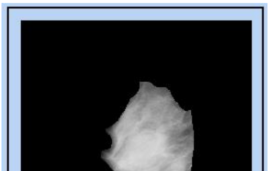
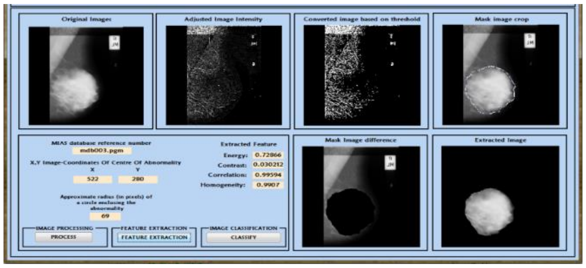


Fig. 7a: Mask ROI image Fig. 7b: Mask crop image Fig.7c: Mask difference image

4.3 Feature extraction

The features are calculated from the ROI to obtain the gray level characteristics which are shape and texture of the lesion and the surrounding tissue. The mask crop image are analyze with the use of function '*texturesGLCM(inputimage)*' to get the texture of the image and obtain the pixel value. This is shown in Figure 8



MIAS database reference number: mdb003.pgm

X,Y Image-Coordinates Of Centre Of Abnormality: X 522, Y 280

Approximate radius (in pixels) of a circle enclosing the abnormality: 69

Extracted Feature

Energy: 0.72866

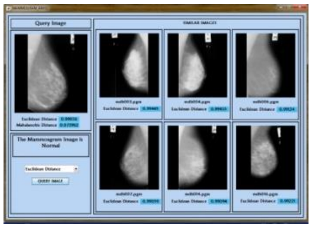
Contrast: 0.030212

Correlation: 0.99594

Homogeneity: 0.9907

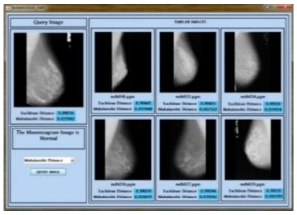
Fig. 8: Extracted region

Matching the feature vectors of the images was carried out by computing the distance (D) between the *hist* of the query image and the i-th database image. The distances (D2) are kept for which, the respective *hist* bins of the query image are larger than a predefined threshold. The sample of a classified image base on Euclidean distance and Mahalanobis distance is shown in Figure 9a and 9b respectively.



The Mammogram Image is Normal

Fig. 9a: Classified image base on Euclidean distance



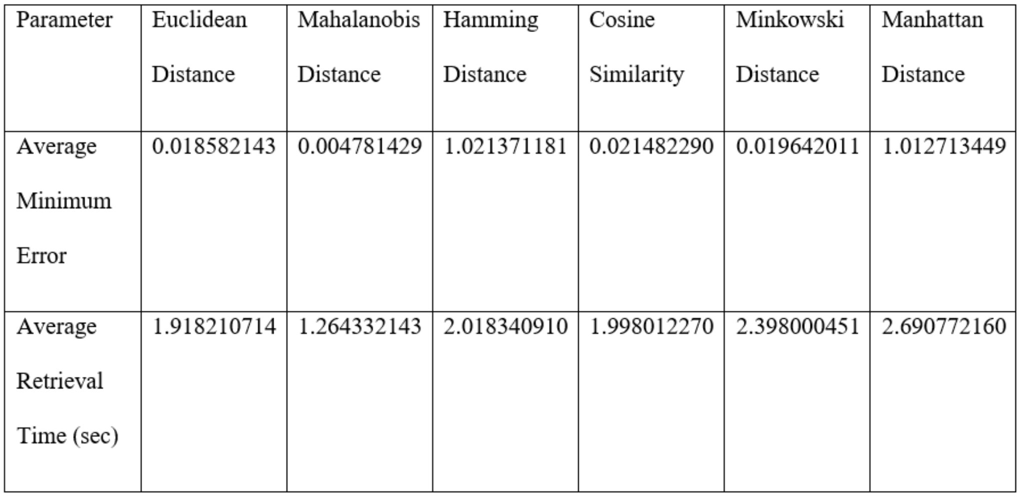
The Mammogram Image is Normal

Fig. 9b: Classified image base on Mahalanobis distance

5. RESULTS AND DISCUSSION

The adequacy of the distance metrics is determined by evaluating whether the image retrieved matches the queried image, by determining the minimum error in the pairs of image and also the time taken to retrieve a similar image. The minimum error is the average distance between the total images retrieved by the system per one query image, when each distance measure has been computed for all the tested images. Retrieval time is the time taken for the system to retrieve all the possible likely images from the image database of one query image. From each of the datasets, a total of 200 samples consisting of both benign and malignant mammograms were arbitrarily queried and the minimum error as well as retrieval time was obtained for each queried image. The average minimum error and average retrieval time was obtained for each of the distance metrics. The summary of the performance of the distance metrics is reported for mini-MIAS dataset in table 1 and BCDR dataset in table 2.

Table 1: Performance comparison result of Distance Measure for mini-MIAS dataset



Parameter	Euclidean Distance	Mahalanobis Distance	Hamming Distance	Cosine Similarity	Minkowski Distance	Manhattan Distance
Average Minimum Error	0.018582143	0.004781429	1.021371181	0.021482290	0.019642011	1.012713449
Average Retrieval Time (sec)	1.918210714	1.264332143	2.018340910	1.998012270	2.398000451	2.690772160

It can be deduced from the results obtained in table 1 that Mahalanobis distance measure outperforms the other distance metrics both in terms of a low minimum error as well as short retrieval time. This means that the time taken to retrieve similar image with the queried one is very short and there is high similarity between the images displayed. This is an indication that this distance metric can be recommended for use in this research domain. Also, the Euclidean distance and Minkowski distance has very little difference in performance in terms of minimum error, these two can be said to have a good performance according to the metrics. The Manhattan distance has the worst performance in terms of retrieval time; hence it can be said to be slow in performance.

Table 2: Performance comparison result of Distance Measure for BCDR dataset

Parameter	Euclidean Distance	Mahalanobis Distance	Hamming Distance	Cosine Similarity	Minkowski Distance	Manhattan Distance
Average Minimum Error	0.011285341	0.002142798	1.031721710	0.012222490	0.019642011	1.044331220
Average Retrieval Time (sec)	1.87120413	1.146214331	2.10183022	1.771080601	2.231000143	1.000128100

Similarly, it can be deduced from the results obtained in table 2 that Mahalanobis distance measure outperforms the other distance metrics in terms of a low minimum error as well as short retrieval time. The Euclidean distance and Cosine similarity has little difference in performance in terms of minimum error. Additionally, all the distance metrics has lesser retrieval time with the BCDR dataset compared to the mini-MIAS dataset. This could be as a result of the images in BCDR been more refined, having pre computed shape, intensity and texture attribute due to capture environment. The implication from this research and considering the datasets benchmarked, Mahalanobis distance measure proof to be a good distance metric in this domain.

Several earlier research in the area of QBIC system employed the use of classification algorithms and the performance are been evaluated using accuracy, sensitivity and specificity. This research attempt to simplify the evaluation of QBIC system using simpler metrics which are easy to compute, and that have not been given much consideration in the literature.

5 CONCLUSIONS

This research compares the performance of six contemporary similarity metrics for retrieval of similar images from a database of mammogram images on QBIC system. Adopting distance measure in the area of QBIC has been sparingly explored, hence the purpose of the study is to determine the performance of image retrieval using GLCM features for each of the similarity measures. The novelty of the methodology in this work is in the art of comparison based on visual features and spatial information realized with the use of GLCM. The GLCM features, which is a second order textural feature extracted from the mammogram images were the feature vectors in which matching was carried out using Hamming, Mahalanobis, Manhattan, Euclidean, Minkowski, and Cosine similarity distance measures. This was done successfully with no specification of upper/lower bound as used by range-based query nor annotation of any form. The empirical result shows that Mahalanobis distance metric outperforms the others in terms of the parameters considered for evaluation which are the retrieval time of similar image and minimum error for the two dataset that was benchmarked for the experiment. This comparison helps the researchers to take decision about a suitable distance measure to adopt for similar experiment in the domain. Future work intends to experiment into extraction and determining the performance of other types of feature vectors such as shapes and fractals as well as higher order feature vectors on the distance metrics. Other evaluation parameters such as precision and recall can be looked into. Ground truth mammogram datasets will equally be explored.

References

[1] Boss R. Chandra, Thangavel K., Daniel A. Pon. (2013): Automatic Mammogram image Breast Region Extraction and Removal of Pectoral Muscle. *International Journal of Scientific & Engineering Research*, Vol. 4, No.5. pp. 1722-1729

[2] Imon Banerjee, Camille Kurtz, Alon Edward Devorah, Bao Do, Daniel L. Rubina, Christopher F. Beaulieu (2018). Relevance feedback for enhancing content-based image retrieval and automatic prediction of semantic image features. *Journal of Biomedical Informatics* 84 pp.123–135

[3] Hong B.W., SohnB.S., (2010). "Segmentation of regions of interest in mammograms in a topographic approach," IEEE Trans on Information Technology in Biomedicine, vol. 14, 1, pp. 129–139.

[4] Menglin Jiang, Shaoting Zhang, ,Hongsheng Li, , and Dimitris N. Metaxas (2015). Computer-Aided Diagnosis of Mammographic Masses Using Scalable Image Retrieval, IEEE Transactions on Biomedical Engineering, VOL. 62, No. 2,

[5] Jun Luo and Hang Kuang, (2009). Content-based image retrieval using combination features. *Computer Engineering and Applications*. Vol. 45 ›› Issue (1): 153-155.DOI: 10.3778/j.issn.1002-8331.2009.01.048

[6] Saxena, P., & Shefali (2018). A Bird's Eye View on Current Scenario of Content Based Image Retrieval Systems. *Journal of Electronics and Communication Engineering*. Vol. 13, No. 3, pp.9-15

[7] Ja-Hwung Su, Wei-Jyun Huang, Philip S. Yu, and Vincent S. Tseng. (2011)."*Efficient Relevance Feedback for Content-Based Image Retrieval by Mining User Navigation Patterns*" IEEE transactions on knowledge and data engineering, vol. 23, no. 3

[8] Niranjan B. Subranian (2019). Available at https://aiaspirant.com/distance-similarity-measures-in-machine-learning/. Accessed May 13th 2019.

[9] Manish Chowdhury, Sudeb Das and Malay Kumar Kundu (2012): Novel CBIR System Based on Ripplet Transform Using Interactive Neuro-Fuzzy Technique. *Electronic Letters on Computer Vision and Image Analysis* vol. 11 no.1.pp.1-13.

[10] Pourghassem, H., & Daneshvar, S. (2013). A framework for medical image retrieval using merging-based classification with dependency probability-based relevance feedback. *Turkish Journal of Electrical Engineering & Computer Sciences.* Vol. 21: pp. 882 – 896

[11] Vipparthi, S.K., & Nagar, S.K. (2014). Color Directional Local Quinary Patterns for Content Based Indexing and Retrieval. *Human-centric Computing and Information Sciences,* Vol.4, pp.1-13.

[12] Bommeswari, B., Siva, K.R. and Karnan, M. (2014) "Computer aided detection algorithm for digital mammogram images – a survey", *International journal of computer trends and technology*, Vol.8, No.3, pp.138-147.

[13] Soni, N., & Pinjarkar, L. (2017). Content Based Image Retrieval (CBIR): review and challenges. *International Journal of Engineering Sciences & Research Technology*, Vol. 6, no.6, pp.171-174. doi:10.5281/zenodo.805409

[14] Mohd. Aquib Ansari, Manish Dixit, Diksha Kurchaniya, Punit Kumar Johari. (2017). An Effective Approach to an Image Retrieval using SVM Classifier. *International Journal of Computer Sciences and Engineering.* Vol. 5, No.6 pp. 62-72

[15] Mutasem K. Alsmadi. (2018). Query-sensitive similarity measure for content-based image retrieval using meta-heuristic algorithm. *Journal of King Saud University– Computer and Information Sciences* Vol. 30. pp. 373–381

[16] Jesi Elizabeth, Govindarajan S., Jayanthi M. (2019). A State of Art on Content Based Image Retrieval systems. *International Journal of Recent Technology and Engineering.* ISSN: 2277-3878, Vol.-8, No-2 S4, pp.297-301

[17] Brij Bhan Singh and Shailendra Patel (2017). Efficient Medical Image Enhancement using CLAHE Enhancement and Wavelet Fusion. International Journal of Computer Applications (0975 – 8887) Vol. 167, No.5. pp. 1-5