Survey on Content Based Image Retrieval for Breast Cancer Detection and Classification from Mammography Masses

Ashebir Dohe Dogoma
Symbiosis Institute of Technology, Department of Computer Science and Engineering
Symbiosis International (Deemed University) Pune, Maharashtra 412115, India.
Email: ashebir.dogoma@sitpune.edu.in

Dipti Kapoor Sarmah
Symbiosis Institute of Technology, Department of Computer Science and Engineering
Symbiosis International (Deemed University) Pune, Maharashtra 412115, India.
Email: dipti.sarmah@sitpune.edu.in*

Abstract

Cancer is one of most known diseases that causes the death of many people in the world. Differently breast cancer is the most common type of cancer which took the life of many women in the last ten years. Various methods and systems are proposed to early detect and diagnose the disease by many researchers and physicians. Content-based mammography retrieval systems are the prominent systems which retrieve feature vectors that are extracted and stored in reference databases in order to compare with query image using similarity measuring methods. The diagnosis and detection of the disease through content-based image retrieval depends on the feedback obtained from the system and experimental knowledge of radiologists. The survey work attempts to overview studies related to previous content-based image retrieval systems and provide an insight into different algorithms used for mammogram segmentation, extraction, classification, challenges, and outcomes of previous works made in the research domain.

Keywords: Breast Cancer, Content Based Image Retrieval, Feature Analysis, Similarity Measurement.

I. INTRODUCTION

Breast Cancer is treated as one of the prominent causes of cancer-related death among women in worldwide. This disease is highly seen in women with age of above 40 but also observed in men. Women as a whole has about 12% chance of having breast cancer, while this accounts for 29% of newly diagnosed cancer in the US [1]. Early stage detection and pre-treatment is highly required as it can substantially reduce the mortality rate. The advancement of studies with the involvement of new technologies targeting breast cancer diagnosis successfully introduced a mechanism to control the disease at early stage. Among those introduced medical research findings, Content-Based Image Retrieval Systems (CBIRS) are effective to minimize the risk of unwanted results of the disease.

CBIRs retrieve feature vector images that are extracted using different extractions methods, which help radiologists to process and analyse the retrieved images based on their contents of shape, texture, colour and intensity unlike other medical imaging processes which depends on high-level information and textual contents of the image [2]. In CBIR, periodic screening of breast images is mostly accepted imaging technique of screening breast cancer [3]. Mammography is the imaging technique for breast screening and describing explored advantages for women older than 40 years [4]. As earlier the cancer diseases are detected, the earlier the treatments can be taken. However, early detection requires exact and potentially reliable diagnosis to discriminate benign, malignant and normal tumours [5]. The diagnosis of lesions from breast image is challenging, resulting in inaccurate detection and classification. This is because of the techniques and algorithms used in various steps of developing CBIR systems [6]. This results in production of a various digital breast images in hospital databases and breast screening centres. To explore such important information in supporting diagnoses, CBIRS need to
Different CBIR systems for breast cancer detection, classifications and detailed methodologies with corresponding algorithms are comparatively highlighted in this survey paper. Furthermore, distinct methods, techniques, and algorithms used in different CBIR systems including performance, accuracy, advantages, disadvantages, limitations and comparative analysis are discussed.

The paper is presented as the following sections, Section I depicts related research studies, Section II deals with description of different image pre-processing and segmentation methods, different approaches of feature extraction, feature selection and classification algorithms used for CBIR systems, conclusion and feature works are presented in Section IV and finally reference are drawn at the end.

II. RELATED WORKS

CBIR is most commonly used for early treatment and diagnose of breast cancer as many research studies explain. In order to enhance the identification and discrimination of breast cancer from mammography, different techniques and algorithms have been used by various researchers as far as different studies are concerned [8]. Digital image (Mammogram) pre-processing, feature selection, feature extraction, classification and result evaluation are common steps of CBIR systems.

Many studies attempted towards breast cancer CBIR in the last few decades, one of the studies was the work of Liyang Wei et al. [9] investigated how retrieved similar cases can be used to enhance the performance of a numerical classifier. Their motive was adaptively combining local proximity resources into a classifier, which in turn led to accurate classification, thereby leading to an enhanced “second opinion” to radiologists. They used support vector machine algorithm for classification and radiologists’ result by similar images retrieval on 200 images of data set. They have evaluated the system performance and achieved classification accuracy from 0.78 to 0.82 precision level. The obtained accuracy was not confidential because of the data set used was small in size which leads to generalization errors. Sergio Francisko et al. [10] studied genetic feature selection approach of increasing the ranking quality of medical image retrieval. The objective of the study was developing and evaluating methods and techniques of selection specific features to enhance the accuracy level of image retrieval systems. Their final result states that the newly proposed and implemented feature selection technique increased the accuracy of matching and importantly reduced the data dimensionality. Using genetic algorithms improved only local search which was the challenges of the study and textual information is additionally used for retrieval process.

Chia-Hung et al. [11] presented systems that retrieve pathologically similar images with the given example from reference image database. Different methods of segmentation based on Breast Imaging Reporting and Data System (BI-RADS) standards, Euclidean and Weighted Euclidean Distance Measures are used for similarity measures. The outcome of the study was reported as round shape from shape type shows effective discrimination and circumscribed margin from margin type can be effectively discriminated. Short response time of the system, difficulties in identifying similar shapes and problem of friendly user interfaces were challenges of the work. Developing and testing interactive system with Computer Aided Diagnosis (CAD) of breast cancer discrimination and a diagnosis was another work related to CBIR by Xingwei Wang et al. [12]. They used a genetic algorithm for optimization, K-nearest neighbour for similarity computing and leave-one-out folding for classifying suspicious regions. Fortunately, the designed CAD system fails to segment the digital medical images and the system was evaluated based on classification only. The evaluated result shows in Cranio-Caudal (CC) view detection and is achieved of 0.96 whereas classification is 0.74 precision. But in Mediolateral-Oblique (MLO) view 0.91 for detection and 0.80 for classification precision are recorded.

Menglin Jiang et al. [13] worked on scalable image retrieval model for classification and retrieval of images. Features extraction are done by Scale-Invariant Feature Transform (SIFT) methods and whereas hierarchical k-means algorithm for similarity measures are applied. They explored the new method for efficient reference database by building vocabulary tree which finally led them to increase number of clusters which in turn results in better precision in both retrieval and classification. According to the authors additional features can be analysed to improve retrieval precision and accuracy. Rahimeh Rouhi, Mehdif Jafari et al. [14] presented an article on benign and malignant breast tumour discrimination using segmentation techniques of region growing and cellular neural networks. The proposed methods and algorithms are tested by two separate datasets from Digital Database for Screening Mammography and Mammographic Image Analysis Society (MIAS). Evaluation of the result was done by algorithms like random forest, Naïve Bayes, Support Vector Machine (SVM) and K-Nearest Neighbour (KNN) which results in promising outcomes. The obtained sensitivity, specificity and accuracy measures were 96.87%, 95.94%, and 96.47% respectively. The usage of varying data sets brought various outcomes which in turn results in non-confidential generalization.

The article by Rozita Rastghalam et al. [15] described breast cancer diagnosis using probable texture feature and decision-level fusion based discrimination using Hidden Markov Model (HMM) on thermography images. Features from the digital images are extracted by asymmetric feature extraction, local binary pattern (LBP), and region of interest (ROI) cropper for segmentation, HMM classification algorithms are used as classifiers. The designed techniques were tested on varying datasets
of thermography images which results in false negative rate of 8.3% and false positive rate of 5%. Asymmetric analysis was the drawbacks of the methods used because of its inaccurate diagnosis.

Breast cancer detection feature examination for diagnosis via jointly sparse learning is an article by Heng Kong et al. [16]. Jointly Sparse Data Analysis (JSDA) methodology was used with its algorithm, obtained result shows that 90% classification accuracy, showing that tumour-size, deg-malign and irradiate are the basic determinant features for identification and classification. Finally, the work revealed that JSDA is the effective compared to other related algorithms. As a drawback, the study was done only to identify the most discriminant key features for further analysis. Dongxiao Gua et al. [17] explored a case-based reasoning diagnosis method which uses weighted different value distance metric. It basically aimed to predict benign and malignant tumours as part of the diseases. They used genetic algorithm and five-fold cross-validation for classification methods. The result obtained depicts that the system was evaluated for persons in use through an interview and the participants approved the usefulness to improve medical quality, in saving medical resource and improving oncologists' specialization level. But, the study was done only to specific features and small datasets were used, they recommend feature work on various key features and with large datasets. Finally, supervised based CBIR approach was proposed by Lazaros Tsochatz et al. [18]. The main motive of the study was integrating CBIR with CAD to support physicians in determination process by considering the outcome of retrieval process on discrimination and margin specific feature discovering. The application of SVM in retrieval processes and implementation of new model for texture descriptor, called Modified Histogram Oriented Gradient (MHOG), designed to discriminate margin and core salient features of breast image. Contour of the masses is drawn by semi-automatic segmentation which is selected and applied with aim of acquiring pretty clear margins and specific regions of the image. Next, the identified mass region are extracted into various shapes, texture-specific regions which later transformed into a new representative vector with the help of selected classifier algorithms. Highest result was achieved in the case of speculated and micro lobulated masses among the identified key features. The speculated is from a region form shape whereas micro lobulated is from margin specific masses.

III. COMPARATIVE ANALYSIS OF VARIOUS METHODS AND ALGORITHMS USED IN CBIR FOR BREAST CANCER DIAGNOSIS AND DISCRIMINATION

The approaches and techniques used in each phases of CBIRs for breast cancer diagnosis and discrimination have its own effect on an accurate retrieval of the query image [19]. The effective the methods are chosen the better accurate outcomes could be obtained. Various feature extraction, feature selection, and classification algorithms used by different researchers and studies [20]. The accurate implementation and successful outcome is gained by selecting powerful algorithms and implementing those methods in effective and efficient manner. The tables below depict the comparative summary of distinctive methods and algorithms used in CBIR for breast cancer diagnosis and discrimination used in various research works. Table 1 describes about summary of image segmentation approaches used in previous studies. The techniques of feature extraction are contrasted in table 2, and lastly table 3 explains about different classification algorithms used in various CBIRS for breast cancer diagnosis.

<table>
<thead>
<tr>
<th>Methods/Algorithm</th>
<th>Type of the method</th>
<th>Strength</th>
<th>Weakness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Threshold</td>
<td>Similarity-Based</td>
<td>a. Simple to implement</td>
<td>a. No guarantee of object coherences</td>
</tr>
<tr>
<td></td>
<td></td>
<td>b. Fast on similar images</td>
<td>b. Sensitive to noise</td>
</tr>
<tr>
<td></td>
<td></td>
<td>c. Good for documented images</td>
<td></td>
</tr>
<tr>
<td>Region growing</td>
<td>Similarity-based</td>
<td>a. Based on growing of initial pixels</td>
<td>a. Arrangement and position of the image results in a group segmentation</td>
</tr>
<tr>
<td></td>
<td></td>
<td>b. Requires prior knowledge</td>
<td>b. regular division leads to over segmentation</td>
</tr>
<tr>
<td></td>
<td></td>
<td>c. Guarantee connected neighbours</td>
<td></td>
</tr>
<tr>
<td>Clustering segmentation</td>
<td>Cluster based</td>
<td>a. clusters pixels with similar characteristics</td>
<td>a. Depends on the metadata of the image</td>
</tr>
<tr>
<td>Edge based segmentation</td>
<td>Discontinuity based</td>
<td>a. Easy to implement</td>
<td>a. not good for noise and edgeless images</td>
</tr>
<tr>
<td></td>
<td></td>
<td>b. Easy to understand</td>
<td>b. Not good smooth</td>
</tr>
</tbody>
</table>

TABLE 1. Summary of some Image segmentation techniques
TABLE 2. Image Feature Extraction methods

<table>
<thead>
<tr>
<th>Methods/Algorithms</th>
<th>Strength</th>
<th>Weakness</th>
<th>Classification Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grey scale co-occurrence matrix</td>
<td>a. Gives excellent result for classification &lt;br&gt;b. Powerful and commonly used</td>
<td>a. Takes high time &lt;br&gt;b. Requires lower special frequency of an image &lt;br&gt;c. Computational cost</td>
<td>1.596 1.89 2.345</td>
</tr>
<tr>
<td>Wavelet transform</td>
<td>a. Represents an image in 2D &lt;br&gt;b. Multi-resolution analysis &lt;br&gt;c. Better time for high frequency &lt;br&gt;d. Recently preferred method</td>
<td>a. Poor frequency resolution for high frequency &lt;br&gt;b. Discrete transform is less efficient &lt;br&gt;c. Requires to invest some energy to get proper result.</td>
<td>100% 100% 98.8%</td>
</tr>
<tr>
<td>Independent component Analysis</td>
<td>a. Represents multidimensional vectors as non-Gaussian vectors &lt;br&gt;b. Similar to PCA</td>
<td>a. Works to independent complex dataset &lt;br&gt;b. Offer most complicated results</td>
<td>-Clearly not known</td>
</tr>
<tr>
<td>Non negative matrix factorization</td>
<td>a. Recently developed method for part founding &lt;br&gt;b. Depends on linear representation of non-negative matrix &lt;br&gt;c. Always additive</td>
<td>a. Allow no subtraction in the processes</td>
<td>-Dependent on other factors</td>
</tr>
</tbody>
</table>

TABLE 3. Classification algorithms comparative analysis

<table>
<thead>
<tr>
<th>Methods/Algorithm</th>
<th>Strength</th>
<th>Weakness</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Support vector machine</td>
<td>a. Offers best classification accuracy on training data &lt;br&gt;b. Renders more efficient classification on future data &lt;br&gt;c. Does not make any strong assumption &lt;br&gt;d. Less over fit the data</td>
<td>a. Very slow in training and testing &lt;br&gt;b. Not preferable for discrete dataset &lt;br&gt;c. High arithmetic and computational complexity</td>
<td></td>
</tr>
<tr>
<td>Random Forest</td>
<td>a. No need of data pruning &lt;br&gt;b. Less overfitting &lt;br&gt;c. Most effective and versatile &lt;br&gt;d. Most robust to noise</td>
<td>a. Hard to develop bad random forest &lt;br&gt;b. May be biased because of categorical labelled variables</td>
<td></td>
</tr>
<tr>
<td>Decision tree</td>
<td>a. Robust to errors &lt;br&gt;b. Best suit in problems where instances are represented &lt;br&gt;c. Can handle missing values &lt;br&gt;d. Saves data representation time</td>
<td>a. Less accurate for large decision tree &lt;br&gt;b. Outcomes are expectation based &lt;br&gt;c. Not fit for continuous variables</td>
<td></td>
</tr>
</tbody>
</table>

IV. CONCLUSION AND FUTURE WORK

This survey has focused on detailed review of applications of CBIRS in breast cancer with an overview of previous CBIR systems in diagnosis of the disease. The techniques, methods, and algorithms used in different CBIR systems, their limitations, future challenges and results are deeply addressed with comparative analysis. CBIR systems are widely used for early stage treatment and diagnosis of breast cancer because of their ability to represent low-level information of an image more than visual and textual representation separately, but there remains technical challenges and limitations that requires further investigation which includes various aspects. In basic CBIR system, various imaging modalities are used, digital mammography is proved to be the most accurate imaging technique for image retrieval systems. Selection of realistic imaging modality, better salient part selection, effective extraction method and accurate classifiers lead to fix the limitations and challenges of the existing CBIR systems. In this study various research works done related to CBIR for breast cancer detection and classification has been explained to emphasize the concept. Here the techniques, methods, and
algorithms used in different stages of CBIR systems, applied by scholars to improve the retrieval efficiency and accuracy are discussed.

In the further study, one can expect integrating rich concepts from different fields with CBIRS specifically using low-level information of the image to enrich the challenges and limitations of the problem domain. More research is to be carried out to combine all the key feature of the image shape, colour, intensity and texture with powerful techniques of feature extraction and good machine learning algorithms to achieve maturity in terms of accuracy, efficiency, usability, and performance.

REFERENCES


