Screening of Mammographic Masses using Vocabulary Tree-Based Image Retrieval

Sasikala. V
Final Year, M.E Computer Science and Engineering
and
Valarmathi. P
Head, Computer Science and Engineering,
Mookambigai College of Engineering, Pudukkottai-622502, Tamil Nadu
Email: sasidan.213@gmail.com

Abstract**—** *Computer-aided diagnosis of masses in mammograms is important to the prevention of breast cancer. Many approaches tackle the problem of diagnosis through content-based image retrieval techniques. However, most of the techniques fall short of scalability in the retrieval stage, and restricted diagnostic accuracy. Scalable method for retrieval and diagnosis of mammographic masses overcome this restriction. Specifically, for a query mammographic region of interest (ROI), scale-invariant feature transform (SIFT) features are extracted and searched in a vocabulary tree, which stores all the quantized features of previously diagnosed mammographic ROIs. In addition, to fully exert the discriminative power of SIFT features, contextual information in the vocabulary tree is employed to refine the weights of tree nodes. The retrieved ROIs are then used to determine whether the query ROI contains a mass. The presented method has excellent scalability due to the low spatial-temporal cost of vocabulary tree. Extensive experiments are conducted on a large dataset of 11 553 ROIs extracted from the digital database for screening mammography.*

Index Terms- *Computer-aided diagnosis (CAD), Content based image retrieval (CBIR), breast cancer*

I. INTRODUCTION

Breast cancer death rates are higher than that of any other cancers for women in the U.S. Approximately 39 520 women in the U.S. died from breast cancer in 2011, although death rates have been decreasing. These decreases are likely the result of treatment advances, increased awareness, and early detection. Screening mammography is one of the most effective techniques for the early detection of breast cancer. A radiologist typically examines a mammogram to check for signs of cancer. Computer-aided detection (CADe) system prompts the radiologist to reexamine the films. When using a CADe system with mammography, a radiologist still reads the mammogram, but a computer program also evaluates the mammogram and highlights suspicious regions for the radiologist to review. Finally, the radiologist identifies true areas of concern before making a final diagnosis. The CADe system in screening mammography serves as a second opinion that calls attention to abnormalities and avoiding unnecessary biopsies. When two radiologists make different diagnoses of a mammogram, the CADe system can provide an objective machine opinion for them to reconsider. Therefore, CADe systems have been developed to assist radiologists and increase the accuracy of diagnosis.

As an alternative solution, some CAD methods utilize content-based image retrieval techniques specifically; they compare the current case with previously diagnosed cases stored in a reference database, and return the most relevant cases along with the likelihood of a mass in the current case. Compared with classification-based approaches, these methods could provide more clinical evidence to assist the diagnosis, and therefore attract more and more attention. For example, template matching based on mutual information was utilized to retrieve mammographic regions of interest (ROIs), and similarity scores between the query ROI and its best matches were used to determine whether it contained a mass. This approach was further studied using more similarity measures (such as normalized mutual information). Features related to shape, edge sharpness and texture were adopted to search for mammographic ROIs with similar masses. For the same purpose, 14 image features and a k- nearest neighbor (k-NN) algorithm were applied in. This method was improved by removing poorly effective ROIs from the reference database. These methods have shown great value of CBIR techniques in retrieval and analysis of mammographic masses. However, they did not consider scalability and were tested on at most 3200 mammographic ROIs. This drawback limited the diagnostic accuracy, since the larger a reference database is, the more likely those relevant cases are found and a correct decision is made.

In this paper, we propose to solve the above problem through a comprehensive and scalable image retrieval framework, which is illustrated in Fig. 1. Specifically, scale-invariant feature transform (SIFT) features extracted from database ROIs are quantized and indexed in a vocabulary tree. To enhance the discriminative power of SIFT features, statistical information about neighbour nodes in the tree is utilized to refine the weights of tree nodes following . Given a query ROI, SIFT features are extracted and searched in the tree to find similar database ROIs. These ROIs along with the similarities to the query ROI are used to determine whether the query contains a mass or not.

The major contribution of this study is threefold. 1) We introduce the vocabulary tree framework to retrieval of mammographic masses, which is among the first few attempts to tackle the large-scale medical image analysis problem. 2) A general vocabulary tree refinement is selected for the specific mammographic mass retrieval task, which improves the retrieval precision and diagnostic accuracy. 3) We build a dataset with 11 553 mammographic ROIs, which is the largest dataset.

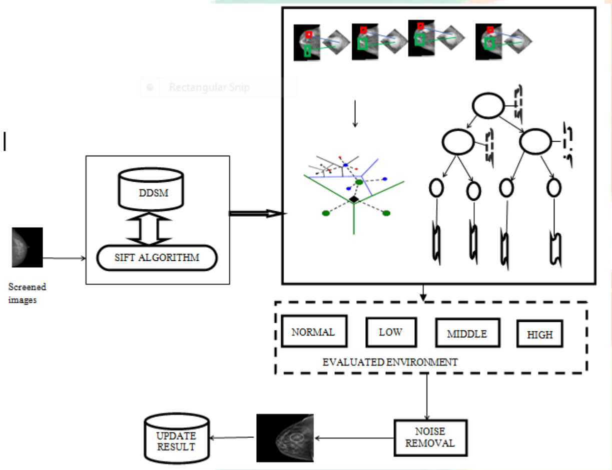


Fig .1. Proposed approach

II. PROPOSED APPROACH

A. SIFT Feature Extraction

Our approach builds upon a popular CBIR framework that indexes local image features using vocabulary tree and inverted files. The local feature we choose here is SIFT. Briefly speaking, SIFT features are extracted in four steps. First, scale-invariant keypoints are detected by finding local extrema in the DoG space. Second, the accurate location and scale of each keypoint are determined using model fitting, and those keypoints with low contrast or poorly localized on an edge are eliminated. Third, for each remaining keypoint, a gradient orientation histogram of its surrounding region at the selected scale is calculated, and the histogram peak is chosen as the keypoint's dominant orientation. Finally, the surrounding region is divided into 4× 4 subregions, an 8-bin histogram of gradient orientations relative to the dominant orientation is computed for each subregion, and all the 16 histograms are concatenated to form a 128-D feature vector. The aforementioned procedure is designed so that the extracted SIFT features are invariant to translation, rotation, scale, a substantial range of affine distortion, viewpoint/ illumination change, and noise addition. SIFT is also very discriminative, i.e., a single feature can be correctly matched from a large database of features. The outstanding robustness and discriminative power catapult SIFT and its variations to the top of local feature performance rankings.

B. Mammogram Retrieval with a Vocabulary Tree

In image retrieval, a straightforward way to match SIFT features would be exhaustive search. Specifically, a query SIFT feature is matched with all the database features, and the database feature with minimum Euclidean distance is identified as the best match. To prune false matches, the second closest database feature is also found, and the ratio of the second-shortest distance to shortest distance, referred to as "uniqueness," can be calculated. Correct matches are expected to have higher uniqueness. However, exhaustive search of SIFT feature is extremely time consuming; therefore it cannot be conducted in large-scale retrieval.

To overcome this problem, we adopt vocabulary tree and Inverted files to quantize and index SIFT features. In this framework, a large set of SIFT features extracted from a separate database are used to train a vocabulary tree through hierarchical *k*-means clustering. The process is illustrated in Fig. 2. Specifically, *k*-means algorithm is first run on the entire training data, defining *k* clusters and their centers. It is then recursively applied to all the clusters, splitting each cluster into *k* subclusters. After *L* recursions, a vocabulary tree of depth *L* and branch factor *k* is built. Each tree node corresponds to a cluster center, and is commonly referred to as "visual word."

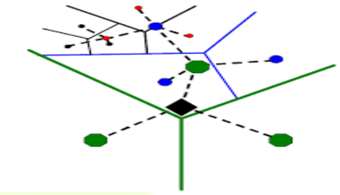


Fig. 2. k-means clustering

Then, all SIFT features extracted from database ROIs are quantized and indexed using this vocabulary tree and inverted files. As shown in Fig. 3., each feature is propagated down the tree by choosing the closest node at each level. Thus, a 128-D SIFT feature is quantized to a 1-D leaf node ID, which represents a path from tree root to leaf. The ID of associated database ROI is then added to the inverted file attached to the leaf node.

Note that an inner tree node also has a virtual inverted file, which is actually a concatenation of all the inverted files attached to its descendant leaf nodes. Unlike a forward file, which lists all the visual words extracted from a ROI, an inverted file records the database ROIs that contain a certain visual word. (The name of inverted files comes from the fact that they are opposite to forward files.) Inverted files significantly outperform forward files with regard to retrieval speed. Given a query image represented as a bag of visual words, querying the forward files would require sequential iteration through each file and to every database feature, therefore, it is technically unrealistic for largescale real applications. On the contrary, searching inverted files only needs to consider those files corresponding to the query visual words, which account for a small portion of all the inverted files. Such advantage is dramatically enhanced with the aid of vocabulary tree, which contains millions of leaf nodes attached with inverted files.

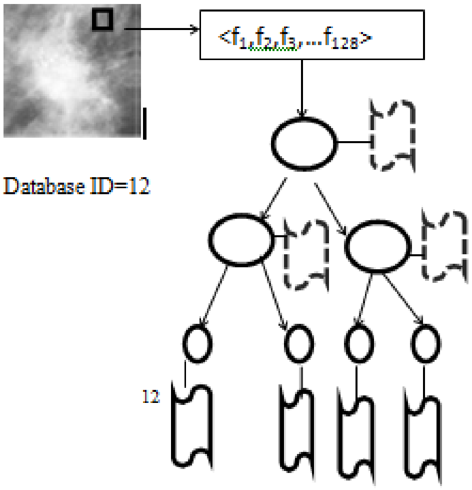


Fig. 3. Quantization and indexing of a database SIFT feature using Vocabulary tree and inverted files

C. Query Image Processing

At last, given a query ROI *q*, SIFT features are extracted and quantized in the aforementioned manner. The similarity score between *q* and a database ROI *d* is calculated based on how similar their paths are. Normally, the tree nodes are weighted using term frequency-inverse document frequency (TF-IDF) scheme or its variations. TF-IDF is widely adopted in vocabulary tree-based CBIR methods. It reflects the importance of a visual word to an image in a collection of images. In brief, TF means the weight of a node is proportional to its frequency in a query ROI, and IDF indicates that the weight is offset by its frequency in all database ROIs.

Formally, *q* is represented by a set of paths q= $\{P_I^q\}^m_{i=1}$, where *m* is the number of features. Each path consists of *L* nodes $P^q_i = \{v^q_{i,l}\}^L_{l=1}$,where $v^q_{i,l}$ denotes the node on the lth level. Similarly, d is represented by d = $\{P^d_j\}^n_{j=1}$, where n is the number of features, and $P^d_j = \{v^d_{j,l}\}^L_{l=1}$, where $v^d_{j,l}$ denotes the node on the lth level. The similarity score between q and d is calculated as the average similarity between all pairs of paths

$$s(q,d) = \frac{1}{m.n}\sum_{i,j} s_P\left(P_i^q, P_j^d\right) \tag{1}$$

where the normalization factor 1/ (*m* · *n*) is used to achieve fairness between database ROIs with few and many features. The similarity between two paths is defined as the weighted count of their common nodes

$$s_p\left(P^q_i, P^d_j\right) = \sum_l w\left(v^q_{i,l}\right).\delta\left(v^q_{i,l}, v^d_{j,l}\right) \tag{2}$$

Where *w* is a weighting function, and *δ* is the Kronecker delta function, i.e., *δ* (*a, b*) = 1 if *a* = *b* and *δ* (*a, b*) = 0, otherwise. In, *w* is defined following the IDF principle as follows:

$$w(v) = idf(v) = \log\frac{N}{N_v} \tag{3}$$

where *N* is the total number of database ROIs and *Nv* is the number of ROIs with at least one path through node *v*. Note that multiple features in *q* quantized to the same node *v* contribute *w* (*v*) multiple times to *s* (*q, d*), which is equivalent to TF. The aforementioned framework allows the use of a very large vocabulary, since its computational cost is logarithmic in the number of visual words. As the vocabulary size increases, leaf nodes become smaller and more discriminative. Therefore, the retrieval precision is improved. In addition, smaller nodes mean that less features from the database need to be considered during similarity calculation. Thus, the retrieval speed is accelerated.

D. Adaptive Weighting of Vocabulary Tree Nodes

The IDF scheme calculates a node's weight based on the whole database, ignoring how frequently it occurs in a specific mammogram. However, generally speaking, features with high frequencies in a mammogram are less informative than those with low frequencies.

Although their IDFs are generally smaller than those of the features extracted from the edge of the mass, they still dominate the similarity score due to large TFs. To avoid such over counting, inspired by descriptor contextual weighting, we incorporate the mammogram-specific node frequencies into IDF scheme to down-weight these features. Suppose the node paths $P^q_i$of query ROI *q* and $P^d_j$of database ROI *d* have the same node $v \in P^q_i \cap P^d_j = \{v^q_{i,l}\}^L_{l=1} \cap \{v^d_{j,l}\}^L_{l=1}$, the node's weight *w* (*v*) in (3) is modified to

$$w^{q,d}_{i,j}(v) = w_p\left(P^q_i\right)\cdot w_p\left(P^d_j\right).\mathrm{idf}(v) \tag{4}$$

where the adaptive weight factors $w_P(P^q_i)$ and $w_P(P^d_j)$ are calculated based on the frequencies of nodes along paths $P^q_i$and P^d_j, respectively. Specifically, let $tf(v^q_{i,l}, q)$ be the TF of $v^q_{i,l}$in *q*, i.e., the number of paths of *q* that pass through node $v^q_{i,l}$, $w_P(P^q_i)$ is defined as

$$w_p\left(P^q_i\right) = \sqrt{\frac{\sum_l w\left(v^q_{i,l}\right)}{\sum_l w\left(v^q_{i,l}\right).tf\left(v^q_{i,l}, \mathrm{q}\right)}} \tag{5}$$

where $w(v^q_{i,l})$ is a weighting coefficient, usually set to $idf(v^q_{i,l})$empirically. $w_P(P^d_j)$ is defined in the same way. The square root in the aforementioned definition is due to the weighting ofboth $w_P(P^q_i)$ and $w_P(P^d_j)$.Note that $w_P(P^q_i)$ is shared for all nodes $v^q_{i,l}$along path P^q_i. In order to determine the importance of a feature P^q_i, $w_P(P^q_i)$ takes into account the features in q quantized to neighbor tree leaves since they also contribute to tf ($v^q_{i,l}$, q). Consequently, nodes in a subtree with more features are heavily downweighted.

III. SYSTEM MODEL

A. Diagnosis of mammographic masses

We consider a computer-aided breast cancer diagnosis system (CABCDS) as shown in Fig. 4. The system contains two modules: context extraction and computer-aided diagnosis.

We consider a sequence of patients numbered t = 1,2,3.... Arrive with a borderline test result. Context extraction module aggregates information x_t from the EHR about a patient t, having a distribution of $f(x_t)$. Then, the computer-aided diagnosis module generates a diagnostic recommendation Rt □ $\{0,1\}$ to the physician, where 0 represents a 6-month imaging follow-up and 1 represents a biopsy. Here, we consider a binary decision, but the approach can easily be extended to incorporate additional choices.

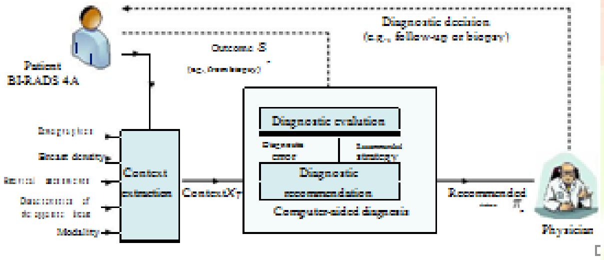


Fig .4. Computer-aided breast cancer diagnosis system model

B. Context Extraction Module

To better assist physicians, the CABCDS system considers a diverse set of contextual information to make sufficiently accurate recommendations. As shown in Table I, the following types of contextual features are considered: patient demographics (e.g., age, race), breast density, assessment history, whether the opposite breast has a high BI-RADS score previously (e.g., achieve BI-RADS 3), and imaging modality (e.g., mammogram, ultrasound).

TABLE I
TYPES OF CONTEXTS AND DESCRIPTIONS

Context	**Description**
Demographics	The characteristics of a patient, age, race, disease history, family medical history, etc.
Breast Density	**Group 1**: The breast is almost entirely fat (fibrous and glandular tissue <25%). **Group 2**: There are scattered fibroglandular densities (fibrous and glandular tissue 25% to 50%). **Group 3**: The breast tissue is heterogeneously dense (fibrous and glandular tissue 50% to 75%). **Group 4**: The breast tissue is extremely dense (fibrous and glandular tissue > 75%).
Historical Assessments	The information contained in previous imaging exam assessments (e.g., whether findings in BI-RADS 3 or higher appear in the past, or whether there is a significant change in the past year).
Characteristic of the opposite Breast	The information of the opposite breast (e.g., whether findings in BI-RADS 3 or higher appear for the opposite breast).
Modality	The modality used for imaging: mammography (MG), ultrasound (US), magnetic resonance imaging (MRI) or computer radiography (CR).

C. Computer-aided Diagnosis Module

This module consists of recommendation generation and diagnostic evaluation steps. The recommendation generation step suggests a diagnostic strategy based on the contextual information and previous diagnostic evaluations. A diagnostic strategy is the approach for selecting an action, either to undergo a biopsy or to follow up, based on the observed contextual information. Given the context x_t of a patient t, $\pi_t(x_t)$ represents the action selected by the diagnostic strategy π_t. The strategy set is denoted by π.

The diagnostic evaluation module collects outcomes of patients. The outcome of the patient t is $s_t(x_t)$, which is either 0 (representing benign) or 1 (representing malignant). If a patient undergoes a biopsy or returns for a short-term followup, the patient's outcome is revealed, where if the patient has been followed up for a certain time and the condition is stable, then the outcome is considered benign. We use $\sigma(x)$ to represent the probability of being malignant for a patient with context x. The evaluation of the diagnostic recommendation is through diagnostic errors. Two types of diagnostic errors are considered: false positive (e.g., if the outcome $s_t(x_t)$ is benign, and the recommended action is to undergo a biopsy) and false negative (e.g., if the outcome $s_t(x_t)$ is malignant, and the recommended action is a short-term follow-up).

D. Diagnostic Recommendation Problem

Based on the given CABCDS system, our design goal is to propose a recommendation algorithm that minimizes the false positive rate (FPR) given a tolerable false negative rate (FNR) η (e.g., < 2%). The trade-off between false positive and false negative rates can be specified by a physician or by an institution. Therefore, the diagnostic recommendation problem is formally written as:

$$\text{minimize FPR} \\ \text{subject to FNR} \leq \eta \qquad (6)$$

IV. EXPERIMENTS

Our experimental dataset is constructed from the digital database for screening mammography (DDSM). DDSM is currently the largest public mammogram database. It is comprised of 2 604 cases, and every case consists of four views, with two views, CC and MLO, for each breast. The masses have diverse shapes, sizes, margins, breast densities as well as patients' races and ages, and are associated with annotations labeled by experienced radiologists. A de identified dataset of 4,640 individuals who underwent screening and diagnostic mammograms at a large academic medical center is used. Patient outcome is derived from biopsy result, which is typically obtained for individuals with a BIRADS score of 4 or 5. Our focus is on analyzing cases that are BI-RADS 4A; this category represents patients whose test results are less suspicious for cancer, raising the concern about unnecessary biopsies. We consider five contextual features, including:(1) patient age, (2) breast density, (3) assessment history (whether or not the immediately preceding exam shows a finding of BI-RADS 3 or above), (4) assessment results for the opposite breast (whether or not the immediately preceding exam shows a finding of BI-RADS 3 or above), and (5) the imaging modality used.

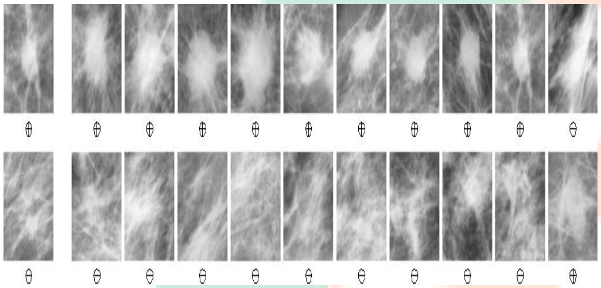
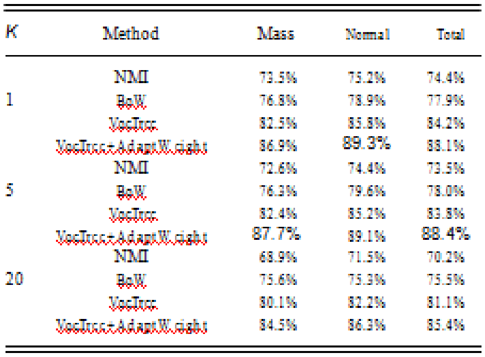


Fig. 7. Two query ROIs (left) and their top K =10 retrieved database ROIs calculated by VocTree+AdaptWeight (right). For each ROI, its class is shown below. Both query ROIs are correctly classified according to a weighted majority vote of their retrieval sets.

First of all, *retrieval precision* is evaluated, which is defined as the percentage of retrieved database ROIs that are relevant to query ROI. Overall the precision changes slightly as the size of retrieval set K increases from 1 to 20. The precisions at top K = 1, 5, and 20 retrievals are summarized in Table II. Two retrieval sets returned by VocTree+AdaptWeight are provided in Fig. 5 for visual evaluation. The results show that our methods, especially VocTree+AdaptWeight, surpass the compared approaches. Detailed results show that many incorrect retrievals are due to the visual similarity between malignant masses and normal ROIs with bright cores and spiculated edges. It is also notable that retrieval precisions for normal regions are generally higher than those for masses. A possible reason is that the database has more normal ROIs than masses, therefore it is easier for a normal query ROI to find similar database ROIs.

TABLE II

RETRIEVAL PRECISION AT DIFFERENT K



K	Method	Mass	Nomal	Total
	NMI	73.5%	75.2%	74.4%
1	BoW	76.8%	78.9%	77.9%
	VocTree	82.5%	85.8%	84.2%
	VocTree+AdaptWeight	86.9%	89.3%	88.1%
	NMI	72.6%	74.4%	73.5%
5	BoW	76.3%	79.6%	78.0%
	VocTree	82.4%	85.2%	83.8%
	VocTree+AdaptWeight	87.7%	89.1%	88.4%
	NMI	68.9%	71.5%	70.2%
20	BoW	75.6%	75.3%	75.5%
	VocTree	80.1%	82.2%	81.1%
	VocTree+AdaptWeight	84.5%	86.3%	85.4%

Second, *classification accuracy* is measured, which refers to the percentage of query ROIs that are correctly classified. The classification accuracies at top K = 1, 5, and 20 retrievals are reported in Table III. Once again, our methods consistently outperform the other two approaches. In addition, the classification accuracy is even better than the retrieval precision, since irrelevant retrievals would not cause a misclassification as long as they remain a minority of the retrieval set. Especially, Voc-Tree+AdaptWeight achieves a classification accuracy as high as 90.8% at K = 5, which is pretty satisfactory.

TABLE III

CLASSIFICATION ACCURACY AT DIFFERENT K

K	Method	Mass	Normal	Total
	NMI	73.5%	75.2%	74.4%
1	BoW	76.8%	78.9%	77.9%
	VocTree	82.5%	85.8%	84.2%
	VocTree+AdaptWeight	86.9%	89.3%	88.1%
	NMI	73.3%	76.1%	74.7%
5	BoW	78.7%	80.3%	79.5%
	VocTree	84.9%	86.7%	85.8%
	VocTree+AdaptWeight	90.1%	91.5%	90.8%
	NMI	71.2%	74.6%	72.9%
20	BoW	77.0%	76.2%	76.6%
	VocTree	81.9%	84.1%	83.0%
	VocTree+AdaptWeight	86.1%	87.7%	86.9%

IV.CONCLUSION

The scalable CBIR is used for the automatic diagnosis of mammographic masses. To retrieve efficiently from a large database, which leads to better retrieval precision and diagnostic accuracy, vocabulary tree framework is employed to hierarchically quantize and index SIFT features. Furthermore, contextual information in the vocabulary tree is incorporated into TF-IDF weighting scheme to improve the discriminative power of tree nodes. A query mammographic ROI is classified using a weighted majority vote of its best matched database ROIs. Extensive experiments are conducted on a dataset including 2 340 mass ROIs and 9 213 CAD generated false positive ROIs, which is the largest dataset to the best of our knowledge.

Excellent results demonstrate our method's retrieval precision, classification accuracy, efficiency, and scalability.

REFERENCES

[1] Chandra PrasetyoUtomo , AanKardiana , Rika Yuliwulandari "Breast Cancer Diagnosis using Artificial Neural Networks with Extreme Learning Techniques", (IJARAI) International Journal of Advanced Research in Artificial Intelligence, Vol. 3, No. 7, 2014

[2] K. Ganesan, U. R. Acharya, C. K. Chua, L. C. Min, K. T. Abraham, and K.-H. Ng, "Computer-aided breast cancer detection using mammograms: A review," *IEEE Rev. Biomed. Eng.*, vol. 6, pp. 77–98, Mar. 2013.

[3] J. Liu, S. Zhang, W. Liu, X. Zhang, and D. N. Metaxas, "Scalable mammogram retrieval using anchor graph hashing," in Proc. IEEE Int. Symp. Biomed. Imaging, 2014, pp. 898–901.

[4] S.-C. Tai, Z.-S. Chen, and W.-T. Tsai, "An automatic mass detection system in mammograms based on complex texture features," IEEE J. Biomed. Health Informat., vol. 18, no. 2, pp. 618–627, Mar. 2014.

[5] Xiaoming Liu, Jinshan Tang, "Mass Classification in Mammograms Using Selected Geometry and Texture Features, and a New SVM-Based Feature Selection Method," IEEE Trans. Med. Imag., vol. 31, no. 6, pp. 1276–1288,
Jun. 2014.

[6] A. Oliver, J. Freixenet, J. Mart´ı, E. P´erez, J. Pont, E. R. E. Denton, and R. Zwiggelaar, "A review of automatic mass detection and segmentation in mammographic images," *Med. Image Anal.*, vol. 14, no. 2, pp. 87–110, 2010.

[7] Y.-H. Chang, L. A. Hardesty, C. M. Hakim, T. S. Chang, B. Zheng, W. F. Good, et al., "Knowledge-based computer-aided detection of masses on digitized mammograms: A preliminary assessment," *Medical physics*, vol. 28, no. 4, pp. 455-461, 2001.