Journal of Biomedical Informatics 44 (2011) 607–614

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Mammogram retrieval through machine learning within BI-RADS standards

Chia-Hung Wei [a], Yue Li [b], Pai Jung Huang [c,d,*]

[a] *Department of Information Management, Ching Yun University, Taiwan*
[b] *College of Software, Nankai University, Tianjin, China*
[c] *Comprehensive Breast Health Center, Taipei Medical University Hospital, Taiwan*
[d] *Graduate Institute of Biomedical Informatics, Taipei Medical University, Taiwan*

ARTICLE INFO

Article history:
Received 13 July 2010
Available online 28 January 2011

Keywords:
Content-based image retrieval
BI-RADS standards
Mammogram
PACS
Image indexing
Feature extraction

ABSTRACT

A content-based mammogram retrieval system can support usual comparisons made on images by physicians, answering similarity queries over images stored in the database. The importance of searching for similar mammograms lies in the fact that physicians usually try to recall similar cases by seeking images that are pathologically similar to a given image. This paper presents a content-based mammogram retrieval system, which employs a query example to search for similar mammograms in the database. In this system the mammographic lesions are interpreted based on their medical characteristics specified in the Breast Imaging Reporting and Data System (BI-RADS) standards. A hierarchical similarity measurement scheme based on a distance weighting function is proposed to model user's perception and maximizes the effectiveness of each feature in a mammographic descriptor. A machine learning approach based on support vector machines and user's relevance feedback is also proposed to analyze the user's information need in order to retrieve target images more accurately. Experimental results demonstrate that the proposed machine learning approach with Radial Basis Function (RBF) kernel function achieves the best performance among all tested ones. Furthermore, the results also show that the proposed learning approach can improve retrieval performance when applied to retrieve mammograms with similar mass and calcification lesions, respectively.

1. Introduction

At hospitals and medical institutions, medical images are produced in ever increasing quantities for diagnostic and therapeutic purposes. As a result, picture archiving and communication systems (PACS) have been developed to integrate imaging modalities and to manage the storage and distribution of images [1]. A crucial requirement in PACS is to provide an efficient search function to access the desired images through the Digital Imaging and Communication in Medicine (DICOM) protocol. Image search in the DICOM protocol is mainly carried out according to the alphanumerical order of textual attributes of image-related information (e.g., such as those found within the DICOM header). However, this requires images to be annotated with text in order to allow images to be accessed by text-based retrieval. As the size of the medical image database grows, it becomes impractical to manually annotate all contents and attributes of the images. The content of images themselves conveys rich information that can be used to search for other images containing similar content. Therefore, content-based image retrieval (CBIR) is expected to integrate into PACS and health database management [2–6]. The importance of searching for similar images comes from the fact that physicians usually try to recall similar cases by seeking images that are pathologically similar to a given image [4].

Breast cancer is the most common cancer among women. In the United Kingdom breast cancer accounts for 30% of all female patients with cancer and approximately 1 in 9 women may suffer from breast cancer sometime during their life [7]. Although breast cancer is a fatal disease, patients still have high chances of survival if malignancy is detected at an early stage. Unfortunately, a high percentage of breast cancer cases are overlooked by radiologists during routine screening [8]. While false negatives can cost lives, false positives can cause panic and lead to unnecessary treatments. It has been reported that only 15–34% of the patients subjected to biopsy are found to actually have malignancies [9]. Mammography is a reliable method for detecting presymptomatic breast cancer [10] and the National Cancer Institute (NCI) recommends that women over the age of 40 and older should have routine screening mammography every 1–2 years [11]. However, the US Preventive Services Task Force (USPSTF) recommends biennial screening mammography for women aged 50–74 years [12]. Even though this change has created a large degree of debate within the medical

* Corresponding author at: Comprehensive Breast Health Center, Taipei Medical University Hospital, Taiwan.

E-mail addresses: rogerwei@cyu.edu.tw (C.-H. Wei), Liyue80@nankai.edu.cn (Y. Li), 0108@h.tmu.edu.tw (P.J. Huang).

1532-0464/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2011.01.012

community, enormous numbers of digital mammograms have been produced at hospitals and breast screening centers. To effectively exploit those valuable resources in aiding diagnoses through PACS networks, content-based mammogram retrieval systems are desirable.

Although mammogram retrieval research has been reported in recent years, some common drawbacks still exist in those existing retrieval systems [13–16]. For instance, the features used in those systems are based on the subjective visual perception of the system designers, rather than on objective definitions of common mammographic standards such as BI-RADS [17]. Hence, those features extracted in those studies should be explained for their medical significance with respect to mammographic lesions and whether those features reflect lesion similarity from radiologists' point of view. In addition, most existing mammogram retrieval systems [13,15,16] do not consider human factors and the semantic gap, which is the difference between descriptions of an object by high-level semantics and representations of low-level pixel data. The semantic gap exists because low-level features are more easily computed in the system design process, but high-level queries are used as the starting point of the retrieval process.

To deal with these issues, a feasible mammogram retrieval system should achieve the following two requirements:

(1) Lesion characteristics should be depicted using the definitions commonly used by medical professionals.
(2) A machine learning mechanism should be embedded in the system for tackling problems arising from human factors and semantic gap.

The Breast Imaging Reporting and Data System (BI-RADS) [17] was developed by the American College of Radiology to enable standardized evaluation of the morphology of breast lesions and categorization of the findings. As the BI-RADS lexicon has been widely used by physicians and radiologists for interpreting mammographic characteristics [18–21], the proposed mammogram retrieval system aims to identify the mammographic lesions based on the definitions specified in the BI-RADS. The definitions of mammographic characteristics used to interpret mammographic abnormalities in this work are provided in [17].

In this work we propose a content-based mammogram retrieval system, which conforms to the BI-RADS standards, for addressing the aforementioned issue. More importantly, the proposed BI-RADS features and the machine learning approach are combined to understand what kind of mammographic lesion the user looks for. The main contributions of this work are summarized below:

(1) A machine learning approach based on support vector machines (SVM) is proposed to exploit the relevance feedback for analyzing the user's information need, thereby seeking for images of greater relevance.
(2) A hierarchical similarity measurement scheme is proposed to automatically assign the weighting for the features at each feature layer.

The rest of this paper is organized as follows. Section 2 provides an overview of the proposed system architecture. Methods for extracting features of mass and calcification conforming to the BI-RADS standards are briefly described in Sections 3 and 4, respectively. Section 5 presents a hierarchical similarity measurement scheme for comparing features. To help understand the user's information need, a SVM-based approach to relevance feedback learning is proposed in Section 6. Performance evaluation of the proposed system is presented in Section 7. Finally, Section 8 concludes this study.

2. An overview of system architecture

The proposed system framework as shown in Fig. 1a can be divided into two components: offline feature extraction for each mammogram in the database and online image retrieval. When new mammograms are added to the image database for the first time, the offline feature extraction component (Fig. 1b) is performed to create a feature descriptor for each mammogram. In the online image retrieval component, the user can submit a query example to the retrieval system to search for desired mammograms. To identify which lesion type the query example belongs to, the system will ask the users to categorize the lesion contained in the example mammogram as either a "mass" or "calcification". The user's categorization can help the system link and utilize the most appropriate feature set (i.e., the mass feature set or the calcification feature set) shown in Fig. 1c. The user intervention avoids using a wrong feature set, directly improving the system performance. The similarities between the feature vector of the query example and that of each mammogram in the feature dataset are computed in a hierarchal manner as shown in Fig. 1d. Finally, the system ranks the similarities and returns the images that are most similar to the query example. This stage is called the query-by-example search for a given query. If the user is unsatisfied with the query-by-example search results, the user can use the relevance feedback function as shown in Fig. 1e to improve the search results. The user provides relevance feedback to the retrieval system in order to refine searches further by tuning the relevance feedback function.

3. Feature extraction of masses

According to the BI-RADS standards, mammographic masses are described by three characteristics – *shape*, *margin*, and *density*. These characteristics are evaluated by medical professionals to determine whether a mass is likely to be benign or malignant. Each mass involves three lesion characteristics, each of which is described by the most similar morphological feature. The combinations of various morphological features from the three lesion characteristics tend to present different degrees of likelihood of being benign or malignant. For instance, masses with round-, oval- and lobular-shaped masses with smooth margins are usually seen as a benign. However, masses of the aforementioned shapes but rugged margins are likely to be malignant. Irregular-shaped masses with indistinct or speculated margin are also likely to be malignant [22]. The methods for extracting the three mass features are described as follows [23]:

- *Shape*: This work uses the Zernike moments for describing mammographic masses because the following reasons: (1) The Zernike basis function satisfies the orthogonal property [24,25], implying that the contribution of each moment coefficient to the underlying image is unique and independent, i.e., no redundant information between the moments; (2) Calculation of Zernike moments do not require knowledge of the precise boundary of an object, i.e., Zernike moments are insensitive to mass segmentation results. This makes Zernike moments suitable for representing masses with obscure boundaries.
- *Margin*: We firstly use Sobel operators [26] to detect the margin of a segmented mass and then the resultant edge map showing the variation in gray levels is obtained to measure its sharpness degree, which is used as a feature to describe margin characteristics. The steps of the proposed method for obtaining margin features are illustrated in Fig. 2.

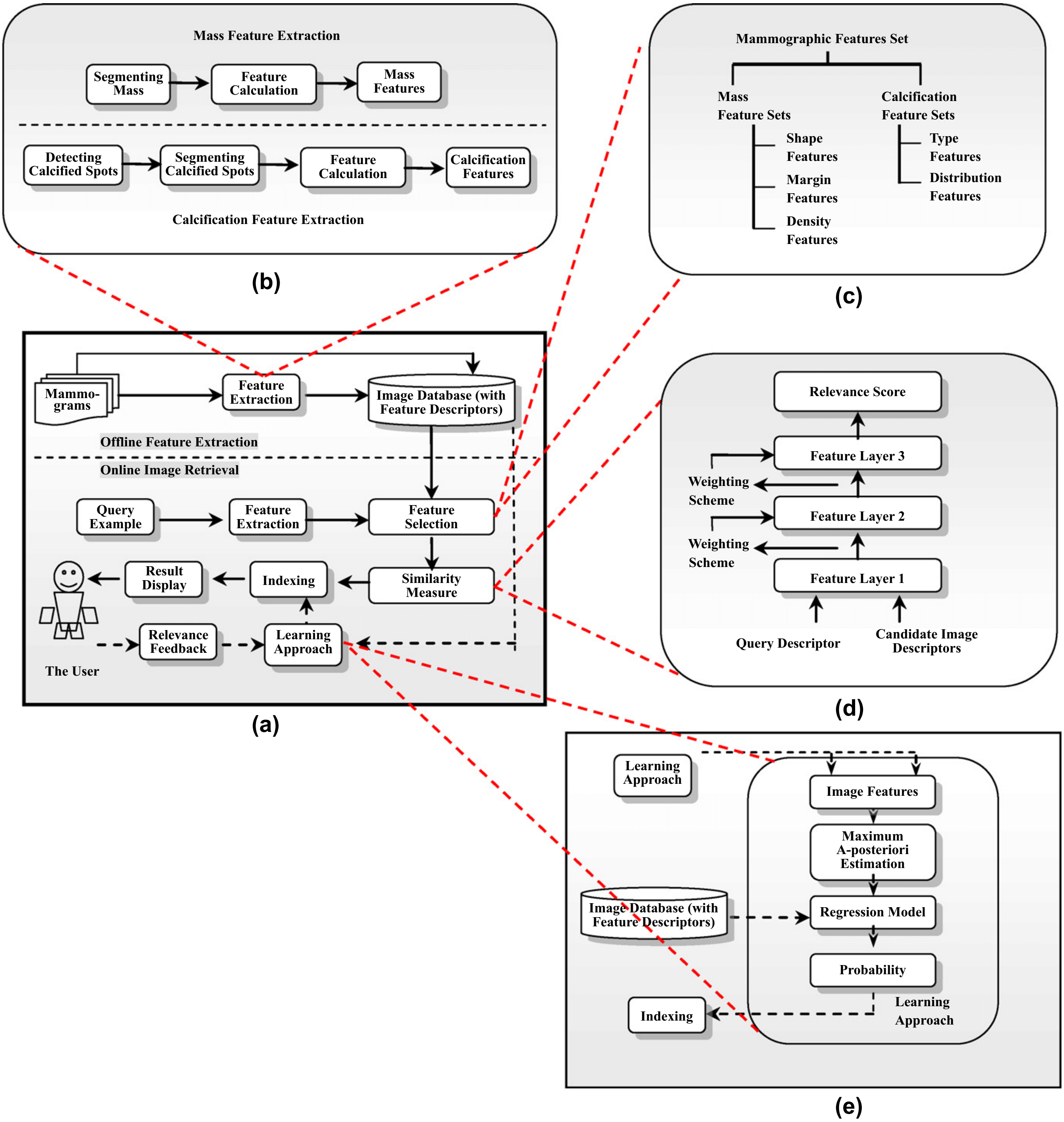

Fig. 1. The proposed system framework and its components for mammogram retrieval. (a) The framework for mammogram retrieval; (b) the feature extraction process; (c) the feature sets; (d) the weighting scheme for similarity measure; (e) the relevance feedback learning approach.

- *Density*: As mammograms are presented in gray level, the density degree is represented by brightness variation of a mass. The steps of the proposed method for calculating mass density feature are indicated as follows: (1) Divide a mass into two regions, the outer and inner regions, where the minor axis of the inner region is approximately half minor axis of the outer region. (2) Calculate the average brightness of the inner and outer regions. (3) Calculate the density degree *DD* for the given mass mammogram according to $DD = \frac{\psi_{\text{inner}}}{\psi_{\text{outer}}}$, where ψ_{inner} and ψ_{outer} are the average brightness of the inner and outer regions, respectively.

4. Feature extraction of calcifications

Before the extraction of calcification features, detection and segmentation of calcified spots in a mammogram are performed to reveal the distributions of calcification and types of individual calcified spots. It is observed that calcifications usually appear as spots which are the brightest objects when compared to other breast tissues. The method proposed in our prior work [23] is adopted to detect and segment calcified spots from mammograms.

According to the BI-RADS definitions, calcification is characterized by type and distribution. Type refers to the characteristics of

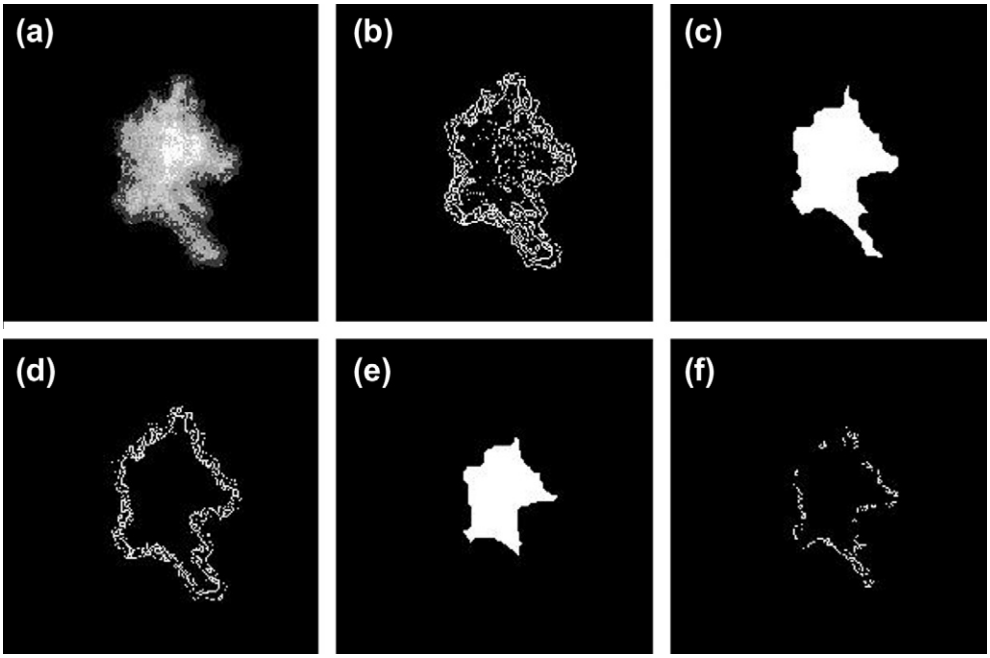

Fig. 2. The margin segmentation process. (a) A segmented mass; (b) gray level variation using Sobel operators; (c) the first shrunken shape is obtained to produce an outer ring; (d) an outer ring is obtained by putting Fig. 3b and c together; (e) a shrunken shape is obtained to produce an inner ring; (f) an inner ring is obtained;

individual calcified spots, such as shape, size, brightness, contrast, margin and density. Distribution refers to characteristics of all the calcified spots in a mammogram considered together, e.g., the range of spread and density of calcification and the arrangement of the calcified spots. As both type and distribution are taken into consideration together, this will result in diverse calcification lesion categories.

4.1. Calcification types

Four features are extracted to describe calcification types as they contain all the information required in the calcification types.

- *Spot Size*: All calcified spots are first found from the segmentation results, and then the average size of all the calcification spots is computed.
- *Spot Shape*: Individual spot shapes are described by the first 15 Zernike moments.
- *Brightness*: This is the average pixel value of all pixels inside the calcified spots.
- *Contrast*: This is the ratio between the average pixel value of calcified spots and that of their surrounding regions, formed by using the morphological 'dilation' operation with a circular structuring element.

In addition, margin and density features are extracted using the method in Section 3.

4.2. Calcification distributions

Calcification distributions in mammograms refer to the characteristics of all the calcified spots in a mammogram considered together. In BI-RADS [17] the calcification distribution are described as "calcifications occupy a small volume of tissue" for a clustered distribution, "arrayed in a line" for a linear distribution, and "scattered in a large volume of breast tissue" or "distributed randomly throughout the breast" for a diffuse distribution. With the above description for calcification distributions, the four distribution features [14] used in this work are as follows:

- *Brightness of Calcifications*: This is the average pixel value of the all pixels inside the calcified spots.
- *Number of Calcified Spots*: When a small number of calcified spots are found in a mammogram, all the detected spots are normally taken into consideration for the distribution characteristic. However, when a large number of spots appear in a mammogram, the significance of each spot in a particular category is reduced. To make approximate comparisons, the number of calcified spots can be represented in generalized orders of magnitude, an estimate measure of quantity expressed as a power of 10. Hence, the number of spots is computed according to $n = 10 \cdot \log_{10}(number)$, where *number* and n are the actual number of spots and the value representing the feature in this work.
- *Dispersion of Calcifications*: Dispersion is defined as a measure of the spatial property of being scattered over an image area, represented by the Zernike moments of the binary image directly transformed from the segmented mammogram. As the segmented mammogram only preserves the calcified spots and removes the underlying breast tissue, the feature extracted from a segmented mammogram directly reflects the calcification distribution.
- *Diffuseness of Calcifications*: This feature, the inter-distance between calcified spots, is computed as follows: (1) A segmented mammogram is transformed into a binary image where the pixels corresponding to the centers of the individual spots are set to 1 and all the rest of the pixels are set to 0; (2) A Delaunay triangulation is next applied to connect the centers of the calcified spots in this binary image; (3) The average mean and standard deviation of the inter-distance between neighboring calcification spots is computed based on this triangulation.

5. Similarity measure

The similarity for any two mammograms is typically measured by calculating the distance between their corresponding feature descriptors (feature vectors) in the feature space. In addition to distance measure, mammogram similarity should further take the following issues into consideration: (1) Importance of individual features: Each feature should be assigned a weight based on its significance on the comparison of lesion similarity; (2) Number of features: Some lesion characteristics are described by one feature value while some are presented by several feature values. Those with more feature values may dominate the similarity measure. Data normalization should be performed before calculating the distance between the two feature vectors; (3) Human vision: When comparing similarity between objects, humans normally examine the outline of objects prior to scrutinizing their inner details. This way of observing objects is applicable to comparisons of mammographic similarity, so features describing the outline of masses and calcifications are more important than those depicting the details.

Due to the aforementioned issues regarding effective comparison of the feature descriptors of query examples and the candidate database images, we propose a similarity measurement scheme which involves the hierarchical arrangement of mammographic features and a weighting distance measure as shown in Fig. 1d. In the hierarchical arrangement, the features are placed at different layers in terms of their importance. Two radiologists were consulted and asked to discuss and come to a consensus on the priority and hierarchical arrangement of features. The features placed at the lower feature layers are regarded as the major characteristics of a mammographic lesion whereas the features at the higher feature layers describe the fine detail of any abnormalities. Table 1 lists all of features used in this system and the layers where the features are located. In mass lesion, the shape feature is placed at the first layer because shape is the primary characteristic used to interpret the outline of masses to human eyes. Furthermore, the shape descriptor, including Zernike moments, may dominate the margin feature and the density feature, each of which includes only one feature value. The margin feature and the density feature are arranged at the second and third layer due to their importance in similarity comparison. The strategy on the arrangement of the mass features is also applied to the arrangement of the calcification features. As calcification mammograms are presented as bright points spread throughout mammograms, the human intuitively overlooks those prominent points in the low resolution and gray level background. Therefore, the number of calcified spots and diffuseness are considered first at layer 1 while dispersion, brightness, and contrast features are placed at layer 2. Because the size of the calcified spots are small and the constraint of image resolution, spot size and shape are not as reliable as those of masses, so these two features (spot size and shape) are placed in layer 3.

The feature arrangement in the similarity measurement hierarchy is designed to meet the following requirements: (1) The measure calculated at a lower layer determines the weighting factor of the next layer, excluding those dissimilar images from similarity calculation in the higher layer; (2) The weighting should be proportional to similarity ranking. To fulfill the afore-mentioned requirements, a weighting distance function is proposed in Eq. (1) to obtain the similarity score E of each feature layer listed in Table 1.

$$E = \sum_{i=1}^{n} w_i v_i \tag{1}$$

where $w_i = (1 - v_{i-1})^k$, $k \in Z^+$, $w_1 = 1$, v_i representing the normalized distance between two descriptors in the range of $[0, 1]$ at the ith layer. Through our previous experiments in [23], $k = 4$ can obtain the best performance. The proposed distance weighting function is utilized for the similarity measure in the hierarchical structure. When the distance between the descriptor of the query image and a database image descriptor at the current feature layer is large, the weighting function will assign a smaller weight factor for the next feature layer, and vice versa.

Table 1
Features extracted for the proposed mammogram retrieval system.

Lesion	Feature layer	Features
Mass	1	Shape
	2	Margin
	3	Density
Calcification	1	Diffuseness, number of calcified spots
	2	Dispersion, brightness, contrast
	3	Spot size and spot shape

6. Interactive retrieval

If the user is unsatisfied with the initial search results, he/she can carry out the interactive search for the same query, as shown in Fig. 1e. In the interactive search process, the task of the CBIR system is to analyze the user's relevance feedback, which is formed from the user's subjective judgement on the returned images, to tune the mammogram feature analysis. Common characteristics in relevant images reveal the user's search target and are what the user is interested in. To analyze the common characteristics and make a prediction, we propose a learning approach to capture the user's intention by analyzing the user's relevance feedback. The proposed approach firstly collects relevance feedback as a set of training data. The set of training data is used to develop an SVM classifier in the real-time retrieval process, which is further applied to conduct probabilistic scaling for the training data. The *a posteriori* probability of membership in the relevant class can be obtained for each candidate image in the database. Note that, in the query-by-example search mode, the similarity between any two images is measured as the Euclidean distance between two points in a multidimensional space, with one dimension representing one feature. During the interactive search stage, image similarity is completely based on the probability estimation. The detailed process for developing the SVM classifier and probabilistic scaling is described in the following sections.

6.1. SVM classifier

Suppose a set X of I retrieved mammograms, X, which have been indicated by the user as *relevant* and *irrelevant*, are given as $\boldsymbol{x} = \{x_i | i = 1, \ldots, I\}$, where x_i is the ith mammogram in x. Let $\mathbf{y} = \{y_i | i = 1, \ldots, I; y_i \in \{1, -1\}\}$ be the class label set with respect to x, with $y_i = 1$ indicating that x_i has been specified by the user as *relevant* and $y_i = -1$ as *irrelevant*. The set of the returned mammograms, x, can be optimally separated by the hyper-plane [27]

$$\boldsymbol{w} \cdot \boldsymbol{x} - b = 0, \tag{2}$$

where $\boldsymbol{w}$ is a normal vector perpendicular to the hyper-plane while b is the displacement of hyper-plane from the original along $\boldsymbol{w}$. The hyper-plane that optimally separates the positive and negative images can be obtained by finding the smallest possible $\boldsymbol{w}$. As the data set $\boldsymbol{x}$ are often not linearly separable, SVM maps the input data into a higher dimensional space through an underlying nonlinear mapping function $\Phi(\cdot)$ and then finds an optimal hyper-plane in the feature space [28].

6.2. Probabilistic scaling

Following the preliminary results, y_i and f_i (i.e., $f_i = \boldsymbol{w} \cdot x_i - b$) are the desired output and the actual output of SVM of data element i, respectively. In the binary class case, the output of the whole training data set is sigmoid, and can be interpreted as the probability of class 1. The logistic likelihood produces the cross-entropy error

$$E = -\sum_{i=1}^{l}[y_i \log f_i + (1 - y_i)\log(1 - f_i)], \tag{3}$$

which represents the negative log likelihood. To apply the output of SVM for logistic regression, y_i is transformed into the probabilistic value t_i with $0 \leqslant t_i \leqslant 1$, which is transformed from

$$t_i = \frac{y_i + 1}{2} \tag{4}$$

The parametric model proposed in [29] can fit the posterior $p_i(y_i = 1|x_i)$. The *a posteriori* probability p_i of the class membership is computed using two parameters λ and η in:

$$p_i = \frac{1}{1 + \exp(\lambda f_i + \eta)} \tag{5}$$

The optimal parameters λ^*, η^* are determined by minimizing the negative log likelihood of the $\boldsymbol{x}$.

$$\min F(t_i, p_i) = -\sum_i t_i \log(p_i) + (1 - t_i)\log(1 - p_i) \tag{6}$$

The problem of finding the optimal parameter set $V^* = [\lambda^*, \eta^*]$ is solved by Newton's method [30]. Newton's method is a numerical optimization method that finds a minimum of a function $F : \Re^n \to \Re^2$ by approaching it with a convergent series of approximations. The search starts in an initial point and computes the step toward the next point. The termination test will be performed for minimization until the minimum is found. Therefore, Eq. (5) can be used to compute the *a posteriori* probability $p_i(y_i = 1|x_i)$ of the class membership for each image in the database.

7. Performance evaluation

7.1. Experiment setup

In this work the data set consists of 1919 mass mammograms and 644 calcification mammograms, obtained from the Digital Database for Screening Mammography (DDSM) [31]. The ground truth of the data set conforming to the BI-RADS specification has been tabulated in Tables 2 and 3 in terms of lesion types, mass characteristics, and calcification characteristics. Tables 2 and 3 indicate pathological characteristic classes, number and percentage of each class, and the abbreviation used to represent the classes selected for performance evaluation in this study. These classes were selected based on two considerations: image quantity and the balance of pathological characteristics used for evaluation [32]. That is, query images were randomly selected from those classes with larger number of images and the pathological characteristics tested belong to different types in mammographic lesions. Therefore, the four mass classes tested include IRR–ILL (265 irregular shape and ill-defined margin mass images), LOB–CIR (178 lobulated shape and circumscribed margin mass images), OVA–OBS (152 oval shape and obscured mass images), and ROU–CIR (59 round shape and circumscribed mass images) classes. The six calcification classes examined are PLE–CLU (255 pleomorphic and clustered calcification images), PLE–SEG (92 pleomorphic and segmental calcification images), PLE–LIN (55 pleomorphic and linear calcification images), AMO–CLU (39 amorphous and clustered calcification images), PUN–CLU (28 punctuate and clustered calcification images), and FIN–LIN (27 fine linear branching and linear calcification images).

Table 2
Ground truth of the mass mammograms used for performance evaluation.

Shape	Margin	Number	Percentage	Abbreviation
Irregular	Ill Defined	265	13.81	IRR–ILL
Lobulated	Circumscribed	178	9.28	LOB–CIR
Oval	Obscured	152	7.92	OVA–OBS
Oval	Circumscribed	134	6.98	
Lobulated	Obscured	105	5.47	
Oval	Ill Defined	102	5.32	
Lobulated	Ill Defined	91	4.74	
Round	Circumscribed	59	3.07	ROU–CIR
Lobulated	Microlobulated	56	2.92	
Round	Obscured	42	2.19	
Oval	Microlobulated	33	1.72	
Round	Ill Defined	33	1.72	
Irregular	Obscured	25	1.30	
Oval	Spiculated	25	1.30	
Miscellaneous		619	32.26	
Total		1919	100.00	

Table 3
Ground truth of the calcification mammograms used for performance evaluation.

Type	Distribution	Number	Percentage	Abbreviation
Pleomorphic	Clustered	255	39.60	PLE–CLU
Pleomorphic	Segmental	92	14.29	PLE–SEG
Pleomorphic	Linear	55	8.54	PLE–LIN
Amorphous	Clustered	39	6.06	AMO–CLU
Punctate	Clustered	28	4.35	PUN–CLU
Fine linear branching	Linear	27	4.19	FIN–LIN
Fine linear branching	Clustered	25	3.88	
Pleomorphic	Regional	19	2.95	
Fine linear branching	Segmental	16	2.48	
Amorphous	Segmental	10	1.55	
Miscellaneous		78	12.11	
Total		644	100.00	

The DDSM database provides the chain codes of the suspicious regions and metadata of each abnormality using the BI-RADS lexicon and contains associated ground truth information about the locations of abnormalities, which were indicated by at least two experienced radiologists [31]. Metadata include the date of study and digitization, the breast density and assessment categories, a subtlety rating, the type of pathology and detailed categorization of the nature of the perceived abnormality. With these chain codes, the outlines of the abnormalities can be identified, enabling us to crop the regions of interest (ROIs) from the full sized mammograms. The evaluation procedure of this study is described as follows. We call the first round of retrieval without user intervention (i.e., relevance feedback) as the Query-By-Example (QBE) mode. After the QBE mode, if the user is unsatisfied but willing to provide relevance feedback for further search (called the Relevance Feedback (RF) mode), the system allows the user to identify an arbitrary number of the returned images as relevant and irrelevant ones. At the end of each round of retrieval (either QBE or RF), the system returns 10 pages of hits with descending similarity rankings, each page containing nine mammograms. Individual observers are likely to provide different relevance feedback due to the fact that relevance for each returned mammogram is subjective. To test the effectiveness of the proposed learning method, we utilized the ground truth of the DDSM database to determine the relevance for each returned mammogram. The results of 10 pages (i.e., 90 mammograms) were identified as either relevant or irrelevant ones, which serve as a set of training data to find the optimal parameters in Eq. (5).

7.2. Results

Precision and recall are the measures commonly used for evaluating the performance of CBIR systems [33]. Precision represents the ratio of the number of relevant images retrieved to the total number images retrieved, while recall is the ratio of the number of relevant images retrieved to the total number relevant images stored in the database. In this work, each precision-recall curve (see Figs. 3–6) represents the performance of one approach and consists of 10 data point, with the *i*th point from the left corresponding to the performance when the first *i* pages of hits are taken into account. A precision-recall curve stretching longer horizontally and staying high in the graph indicates that the corresponding algorithm performs relatively better.

To verify the effectiveness of the proposed system, we first performed tests using different mass and calcification lesions at the QBE mode. Fig. 3 demonstrates that the system performs with the highest precision rates when the query mammograms contain the ROU–CIR (round-shape and circumscribed-margin) masses. This result indicates that the round shape is relatively easier to depict than other shapes. The recall rates for retrieving mammograms with IRR–ILL are relatively higher due to the small number of mammograms with IRR–ILL. In Fig. 4, the three pleomorphic related classes are associated with higher precision rates than the other three non-pleomorphic classes in QBE mode, indicating that the extracted features are sensitive to the characteristic of pleomorphic spots. The result also shows that some PLE–SEG and PLE–CLU images can be identified and ranked at the very top. Their precision rates both decrease rapidly at the 2nd and 3rd pages, and then become stable, thereby resulting in the sudden drops in the curves.

In addition to retrieving in the QBE mode, we also tested the effectiveness of the retrieval system with the relevance feedback (RF) mode. The proposed SVM learning approach was examined and evaluated for retrieval with different kernel functions and their parametric settings as follows: Radial Basis Function (RBF) ($\sigma = 2$, $C = 100$), Polynomial ($p = 2$, $C = 100$), and Splines ($k = 1$, $C = 100$). Another two approaches to relevance feedback learning, namely logistic regression and query point movement, are also tested for comparison. Logistic regression (LR) is a mathematical modeling approach that is used to describe the relationship of real-valued independent variables to a dichotomously dependent variable [34]. The idea of query point movement (QPM) is adopted to move the point of the refined query toward the region in the feature space that contains the relevant images specified by the user. The result shown in Fig. 5 compares the aforementioned algorithms for relevance feedback learning. The precision-recall curves demonstrate that the SVM algorithm with RBF kernel function achieves the best performance among all tested kernels. Furthermore, the proposed SVM algorithms with three different kernel functions present higher precision than that of LR and QPM. The results indicate that the proposed SVM algorithm, no matter which kernel function is associated with, is indeed superior to other relevance feedback learning.

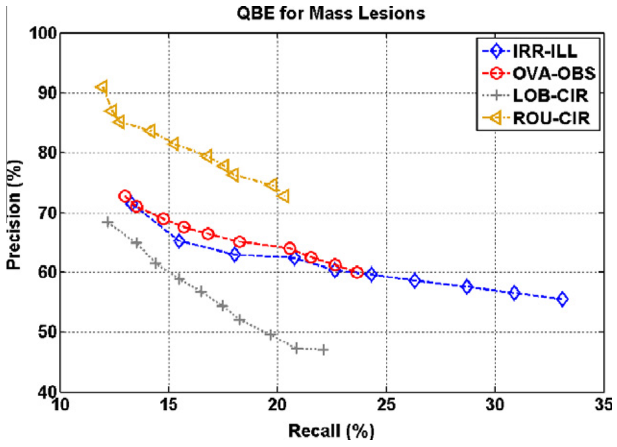

Fig. 3. Evaluation of mammogram retrieval on different mass lesions when considering both shape and margin characteristics.

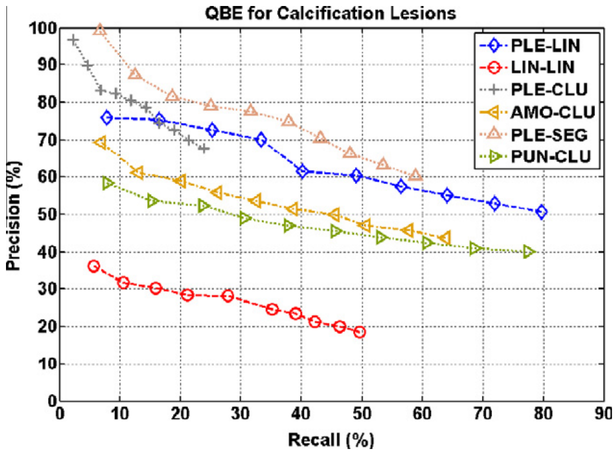

Fig. 4. Evaluation of mammogram retrieval on different calcification lesions when considering both type and distribution characteristics.

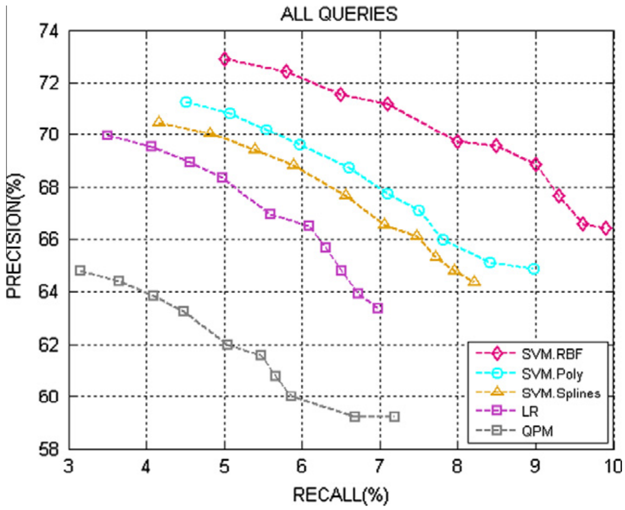

Fig. 5. A comparison of algorithms for relevance feedback learning.

7.3. Discussion

The result shown in Fig. 6 is intended to compare retrieval performance between the QBE mode and the RF mode, i.e., we compared the performance with and without involving RF after the QBE round of search. In the QBE mode, query mammograms are submitted to seek mammograms similar to the query example, while in a RF mode, relevance feedback is provided to refine the current search results by applying the proposed SVM learning approach. The SVM approach uses a SVM classifier with the RBF kernel to find the hyper-plane, which separates positive and negative examples.

The precision-recall curves in Fig. 6 demonstrate that, as we compare QBE's and RF's curves, the proposed learning approach improves the accuracy by as high as around 72% and 74%. It is also worth noting that Calc.QBE is slightly higher than Mass.QBE, which reflects the fact that the effectiveness of feature extraction for calcification lesions is better than that for mass lesions. As the distribution of calcifications consists of a group of calcified spots, the calcification features can tolerate the potential imprecision and some misses in spot detection and the noises of feature extraction. As compared with the calcification features, the features of mass lesions are relatively more sensitive to any imprecision in segmentation and feature extraction of mass lesions.

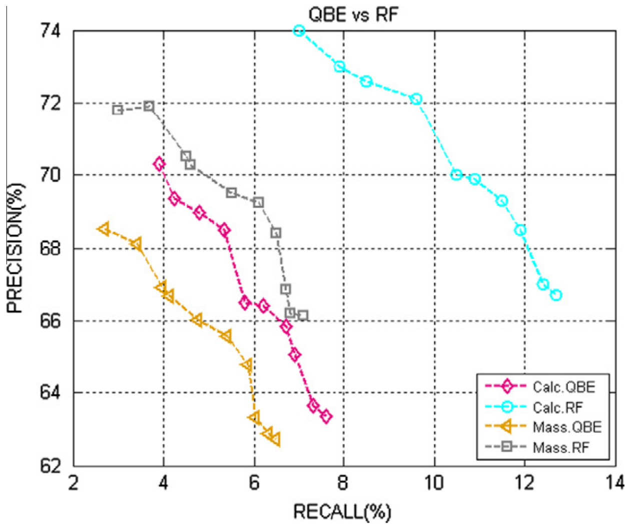

Fig. 6. A comparison of retrieval performance between the QBE mode and RF mode.

8. Conclusions

The contribution of this paper is threefold. Firstly, we have proposed a set of mammogram descriptors according to the BI-RADS definitions. Secondly, we have also presented a novel content-based image retrieval system with a relevance feedback loop that enables the user to bridge the semantic gap between high level human subjective perception and low level machine vision by interacting with the retrieval system. Thirdly, we presented a hierarchical similarity measurement scheme which automatically assigns weighting factors to the more reliable features/descriptors and allows the calculation of the weighting factors of the less reliable features to be conditioned on the more reliable ones. In the future we will explore long-term learning in relevance feedback. Since the relevance feedback approach used in the system does not accumulate any relevance feedback information for use in different sessions, even if the user searches for a specific image he or she reviewed before, the user still must go through the same relevance feedback process to find that image. Therefore, a long-term learning algorithm is required for a real-time system in order to accumulate the user's search information and utilize it to shorten the retrieval time and the relevance feedback process during future query sessions. In addition to the retrieval effectiveness, the usability of the human–computer interaction should be evaluated for the retrieval system. Usability attributes evaluated for the user interface include efficiency, memorability, errors and satisfaction.

References

[1] Huang HK. PACS, image management, and imaging informatics. In: Feng W, Siu C, Zhang HJ, editors. Multimedia information retrieval and management: technological fundamentals and applications. New York: Springer; 2003. p. 347–65.

[2] Bray B, Brown N, Mori AR, Spackman KA, Golichowsky A, Jones RH, et al. Image acquisition context: procedure description attributes for clinically relevant indexing and selective retrieval of biomedical images. J Am Med Inform Assoc 1999;6(1):61–75.

[3] Lehmann TM, Wein BB, Greenspan H. Integration of content-based image retrieval to picture archiving and communication systems. In: Proceedings of medical informatics Europe; 2003.

[4] Traina Jr C, Traina AJM, Araujo MRB, Bueno JM, Chino FJT, Razente H, et al. Using an image-extended relational database to support content-based image retrieval in a PACS. Comput Methods Programs Biomed 2005;80(1):71–83.

[5] Tagare HD, Jaffe C, Duncan J. Medical image databases: a content-based retrieval approach. J Am Med Inform Assoc 1997;4(3):184–98.

[6] Sinha U, Bui A, Taira R, Dionisio J, Morioka C, Johnson D, et al. A review of medical imaging informatics. Ann NY Acad Sci 2002;980:168–97.

[7] Beaver K, Witham G. Information needs of the informal carers of women treated for breast cancer. Eur J Oncol Nurs 2007;11(1):16–25.

[8] Moran S, Warren-Forward H. A retrospective pilot study of the performance of mammographers in interpreting screening mammograms. The Radiographer 2010;57(1):12–9.

[9] Knutzen AM, Gisvold JJ. Likelihood of malignant disease for various categories of mammographically detected, nonpalpable breast lesions. Mayo Clin Proc 1993;68(5):454–60.

[10] Highnam R, Brady M. Mammographic image analysis. London: Kluwer Academic Publishers; 1999.

[11] National Cancer Institute, Mammograms; 2010. <http://www.cancer.gov/cancertopics/factsheet/Detection/mammograms>.

[12] US Preventive Services Task Force. Screening for breast cancer: US preventive services task force recommendation statement. Ann Intern Med 2009;151(10):716–26.

[13] Alto H, Rangayyan RM, Desautels JEL. Content-based retrieval and analysis of mammographic masses. J Electron Imaging 2005;14(2):1–17.

[14] El-Naqa I, Yang Y, Galatsanos NP, Nishikawa RM, Wernick MN. A similarity learning approach to content-based image retrieval: application to digital mammography. IEEE Trans Med Imaging 2004;23(10):1233–44.

[15] Qi H, Snyder WE. Content-based image retrieval in picture archiving and communications systems. J Digit Imaging 1999;12(2):81–2.

[16] Kinoshita S, Azevedo-Marques P, Pereira R, Rodrigues J, Rangayyan R. Content-based retrieval of mammograms using visual features related to breast density patterns. J Digit Imaging 2007;20(2):172–90.

[17] American College of Radiology. The ACR Breast Imaging Reporting and Data System (BI-RADS), 4th ed. Reston (VA): American College of Radiology; 2003.

[18] Baker JA, Kornguth PJ, Floyd Jr CE. Breast imaging reporting and data system standardized mammography lexicon: observer variability in lesion description. Am J Roentgenol 1996;166(4):773–8.

[19] Berg WA, Campassi C, Langenberg P, Sexton MJ. Breast imaging reporting and data system: inter- and intraobserver variability in feature analysis and final assessment. Am J Roentgenol 2000;174:1769–77.

[20] Muhimmah I, Oliver A, Denton ERE, Pont J, Perez E, Zwiggelaar R. Comparison between Wolfe, Boyd, BI-RADS and Tabar based mammographic risk assessment. In: Proceedings of the 8th international workshop on digital mammography; 2006. p. 407–15.

[21] Sampat MP, Whitman GJ, Stephens TW, Broemeling LD, Heger NA, Bovik AC, et al. The reliability of measuring physical characteristics of spiculated masses on mammography. Brit J Radiol 2006;79:S134–40.

[22] Varela C, Timp S, Karssemeijer N. Use of border information in the classification of mammographic masses. Phys Med Biol 2006;51(2):425–41.

[23] Wei C-H, Li C-T, Li Y. Content-based retrieval for mammograms. In: Ma ZM, editor. Artificial intelligence for maximizing content-based image retrieval. Hershey (PA, USA): Idea Group Publishing; 2008. p. 313–39.

[24] Papakostas GA, Boutalis YS, Karras DA, Mertzios BG. A new class of Zernike moments for computer vision applications. Inform Sci 2007;177(13):2802–19.

[25] Wee C-Y, Paramesran R. On the computational aspects of Zernike moments. Image Vision Comput 2007;25(6):967–80.

[26] Gonzalez RC, Woods RE. Digital image processing. Upper Saddle River (NJ): Prentice Hall; 2002.

[27] Burges CJ. A tutorial on support vector machines for pattern recognition. Knowl Discov Data Mining 1998;2(2):121–67.

[28] Scholkopf B, Mika S, Burges CJC, Knirsch P, Muller K-R, Ratsch G, et al. Input space vs. feature space in kernel-based methods. IEEE Trans Neural Networks 1999;10(5):1000–17.

[29] Platt JC. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In: Smola A, Bartlett P, Scholkopf B, Schuurmans D, editors. Advance in large margin classifiers. Cambridge (MA, USA): MIT Press; 1999. p. 61–74.

[30] Rao SS. Applied numerical methods for engineers and scientists. Upper Saddle River (NJ): Prentice Hall; 2002.

[31] Heath M, Bowyer K, Kopans D, Moore R, Kegelmeyer WP. The digital database for screening mammography. In: Proceedings of the fifth international workshop on digital mammography; 2001. p. 212–8.

[32] Xu X, Lee D-J, Antani SK, Long LR, Archibald JK. Using relevance feedback with short-term memory for content-based spine X-ray image retrieval. Neurocomputing 2009;72(10–12):2259–69.

[33] Müller H, Müller W, McG D, Squire S, Marchand-Maillet T, Pun. Performance evaluation in content-based image retrieval: overview and proposals. Pattern Recognit Lett 2001;22(5):593–601.

[34] Kleinbaum DG. Logistic regression. New York (USA): Springer-Verlag; 2002.