Computer-Aided Diagnosis of Mammographic Masses Using Scalable Image Retrieval

Menglin Jiang, Shaoting Zhang*, *Member, IEEE*, Hongsheng Li, *Member, IEEE*, and Dimitris N. Metaxas, *Senior Member, IEEE*

***Abstract*—Computer-aided diagnosis of masses in mammograms is important to the prevention of breast cancer. Many approaches tackle this problem through content-based image retrieval techniques. However, most of them fall short of scalability in the retrieval stage, and their diagnostic accuracy is, therefore, restricted. To overcome this drawback, we propose a scalable method for retrieval and diagnosis of mammographic masses. 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, which demonstrate the accuracy and scalability of our approach.**

***Index Terms*—Breast masses, computer-aided diagnosis (CAD), content-based image retrieval (CBIR), mammography.**

I. Introduction

FOR years, breast cancer remains the second leading cause of cancer-related death among women [1]. Nevertheless, early diagnosis could improve the chances of recovery dramatically: the five-year relative survival rate rises from 24% when breast cancer is diagnosed at distant stage to 99% if it is diagnosed at localized stage [2]. Currently, among all the imaging techniques for breast examination, mammography is the most effective and the only widely accepted method, and it is recognized as a gold standard for breast cancer detection by the American Cancer Society (ACS) [1].

Manuscript received April 15, 2014; revised September 26, 2014; accepted October 24, 2014. Date of publication October 28, 2014; date of current version January 16, 2015. This work was supported by National Science Foundation under Grant NSF-MRI-1229628, Oak Ridge Associated Universities, National Natural Science Foundation of China under Grant 61301269, Sichuan Provincial Key Technology Research and Development Program under Grant 2014GZX0009, and China Postdoctoral Science Foundation under Grant 2014M552339. *Asterisk indicates corresponding author.*

M. Jiang and D. N. Metaxas are with the Department of Computer Science, Rutgers University, Piscataway, NJ 08854 USA (e-mail: menglin.jiang@cs.rutgers.edu; dnm@cs.rutgers.edu).

*S. Zhang is with the Department of Computer Science, University of North Carolina at Charlotte, Charlotte, NC 28223 USA (e-mail: shaoting@cs.rutgers.edu).

H. Li is with the Department of Information Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China (e-mail: lihongsheng@gmail.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TBME.2014.2365494

The major indicators of breast cancer are masses and microcalcifications. Generally speaking, the detection of mammographic masses is even more challenging than that of microcalcifications, since masses have large variation in shape, margin, size, and are often indistinguishable from surrounding tissue [3], [4]. Moreover, even experienced radiologists have substantial interobserver and intraobserver variability in their interpretation of mammograms [5]. Besides, they are often overwhelmed by the enormous mammogram volume generated in widespread screening [6]. Consequently, a considerable portion of retrospectively visible masses is missed by radiologists, and biopsies are frequently conducted on normal tissues [7].

Due to the clinical significance and great challenge of mammographic mass detection, numerous computer-aided diagnosis (CAD) methods have been proposed to facilitate this procedure since the 1960s [8]. A majority of these approaches first segment a query mammogram into several regions, then extract certain features from each region, and finally, classify these regions as mass or normal tissue using the extracted features and pretrained classifiers [3], [4], [6], [9], [10]. For example, Tai *et al.* [11] first segmented adaptive regions of interest (ROIs) as suspicious areas, and then, classified each ROI using complex texture features and stepwise linear discriminant analysis. However, these classifier-based methods are likely to miss masses of "uncommon" appearance or sizes, since it is very difficult for classifiers to model all the training masses. Besides, their performance may be affected by the obscure boundaries of masses, since many of them need to perform image segmentation before mass detection.

During the past decade, content-based image retrieval (CBIR) techniques have gradually gained their popularity among CAD methods for mammograms as well as other medical images. CBIR addresses the problem of searching query images from an image database using visual content inherent in the images [12]–[16], as opposed to text-based image retrieval that utilizes manually annotated keywords. Typically, certain visual characteristics referred to as features are extracted from database images and usually organized in an index structure. Then, for each user-specified query image, the same feature is extracted, and similarities between query feature and database features are calculated with the aid of index. At last, those database images with highest similarities, referred to as retrieval set, are presented to the user.

Mammograms are expected to be an ideal application of CBIR techniques [17], [18], since they depict a limited number of objects and have standard interpretation schemes, such as the breast imaging reporting and data system (BI-RADS) [19]. Specifically, CBIR-based CAD methods first prompt radiologists to

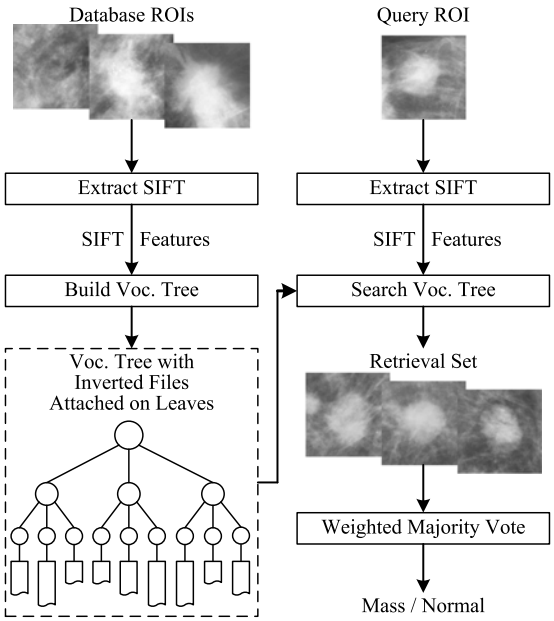

Fig. 1. Overview of the proposed approach.

label an ROI in the query case, then compare it with database ROIs extracted from previously diagnosed cases, and finally, return the most similar cases along with the likelihood of a mass in the query case. Such approaches have several advantages over classifier-based methods. First of all, they could detect unusual masses as long as there are several similar database ROIs. Second, the obscure mass boundary problem is eliminated, since no segmentation is required. Third, they provide more clinical evidence to assist the diagnosis. Last but not least, they can also help improve the performance of picture archiving and communication systems (PACS), augment teaching quality in medical schools, and facilitate radiologist training.

The existing CAD methods have shown great value of CBIR techniques in retrieval and analysis of medical images. However, a vast majority of them fall short of scalability. Instead of utilizing indexing schemes, they compare a query image with at least a considerable portion of database images, making the processing time linear to the total size of database. Consequently, the current mammogram retrieval methods are tested on at most thousands of mammographic ROIs. In contrast, it takes about 2.5 petabytes to store all the mammograms generated in U.S. each year, and all the medical images are estimated to reach 30% of the overall data storage in the world [20]. Apparently, lack of scalability would hamper the utilization of these valuable medical images. On the one hand, it limits the diagnostic accuracy of CAD applications, since the larger a database is, the more likely that relevant cases are found and a correct decision is made [21], [22]. On the other hand, it is infeasible for a practical PACS to retrieve medical images using these techniques. As a result, a scalable CBIR technique has become one of the most urgent problems in medical imaging [23].

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 neighbor nodes in the tree is utilized to refine the weights of tree nodes following [24]. 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. Preliminary results have been published in [25]. Compared with [25], this paper has undergone significant changes. First of all, it is considerably extended to provide more details about our method as well as the techniques at the base of it. Second, the dataset is rebuilt, which now includes 2 340 mass ROIs and 9 213 CAD-generated false positives. Finally, the experiments are substantially improved by adding two compared methods, more evaluation metrics, and a discussion of parameters.

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 [24] 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 to our best knowledge and will be released to public soon. Thorough experiments are conducted on this dataset, demonstrating the efficacy of the presented approach.

The rest of this paper is organized as follows. Section II reviews some relevant work on general CBIR and CBIR-based CAD. Section III describes the proposed approach. Section IV presents the experimental results. Finally, Section V draws a conclusion.

II. Related Work

A. General CBIR Methods

The retrieval accuracy and efficiency of a CBIR method rely heavily on the adopted visual feature. A good feature should obtain a tradeoff between robustness to intraclass variance and discriminability to interclass difference, as well as efficiency of calculation and comparison. Visual features may describe various properties of either a whole image or a local image region, which are usually known as local features and global features. Frequently utilized properties include color, texture, shape, and spatial relationship. Among the numerous features, a local feature named SIFT [26] stands out attributed to its excellent robustness and discriminative power [27]. High-dimensional local features such as SIFT are often quantized for fast retrieval. A quantized local feature is referred to as a "visual word," which is an analogue of "word" in text retrieval, and an image is characterized by a "bag of words" (BoW) [28]. A BoW can be further represented as a histogram, which is regarded as a global feature during indexing and similarity measure.

Another crucial factor to retrieval performance is indexing scheme. In practice, it is infeasible to conduct exhaustive search, which computes a similarity score between the query image and each database image. To solve this problem, an index should be incorporated to narrow down the database images/features

need to be considered during a search. Of all the index schemes, inverted files and hash tables are the most widely used ones for local and global features, respectively. In particular, visual words (quantized local features) extracted from database images are stored in inverted files, which list all the database images per word. During each search, only those files corresponding to query visual words need to be considered [24], [28], [29]. Global features extracted from the database could be indexed in hash tables, where similar features are highly likely to fall into the same bucket in each table, and a query feature is only compared with those database features in its own buckets [30], [31].

Besides visual features and indexing schemes, other techniques such as feature dimension reduction, feature fusion [32], [33] and user interaction [34] can also contribute to CBIR performance. Comprehensive surveys on general CBIR techniques can be found in [12]–[16].

B. CBIR-Based CAD Approaches

The past decade has witnessed many CBIR-based mammographic CAD methods. For instance, template matching is utilized to retrieve similar mammographic ROIs, which are then used to determine whether the query contains a mass [35], [36]. This approach is accelerated by restricting template matching to those database ROIs that share similar entropy values with the query ROI [22]. In order to find similar mammographic masses, features related to texture, shape, and edge sharpness are adopted in [37], intensity, texture, and shape features are fused in [38]. For better visual similarity, users are prompted to rate the margin spiculation of query ROI, and the system only searches from database ROIs with similar spiculation levels [39]. This study is further improved by removing poorly effective ROIs from the database [21]. Several works try to find mammographic masses with similar BI-RADS characteristics [19], such as shape, margin, and pathology. For example, intensity, shape, and texture features are combined using adaptive weights, and user interaction is exploited to optimize the retrieval set [40]. Shape and texture features from two views [cranio-caudal (CC) and mediolateraloblique (MLO)] are fused together, and the retrieved masses are then used to annotate the query mass [41]. A mass ROI is first curvelet transformed, and then, characterized by its marginal curvelet subband distribution [42]. To find similar microcalcification clusters that are consistent with human perception, a similarity learning scheme is proposed to predict radiologists' observations [18]. Features related to intensity, texture, shape, and granulometric measures are employed to retrieve mammograms with similar tissue composition [43]. Recently, Liu *et al.* [44] introduced hashing-based scalable image retrieval to diagnosis of mammographic masses. Specifically, anchor graph hashing (AGH) [31] is employed to compress two features, histogram of SIFT BoW and a global feature named GIST [45], into compact binary codes, and similarity search is performed in Hamming space.

CBIR techniques have also been applied to other medical images and videos. Relevant positron emission tomography (PET) images of human brains are retrieved using a physiological kinetic feature [46]. This method is further extended to deal with 4-D dynamic PET images, which are first segmented into volumes of interest, and then, retrieved using visual, functional, and textual features [47]. Computed tomography (CT) images of chest are first classified to one disease category, and then, searched for similar images using features corresponding to this disease, which are automatically chosen from 125 features related to intensity, texture, and geometric properties [48]. Similarly, a subset of more than 300 features related to intensity, texture, morphology, and spatial relationship is selected to retrieve images of lymphoma cells [49]. Several methods adopt the BoW framework [28] to quantize local features like SIFT. For instance, each endomicroscopy video is represented as a BoW of "dense" SIFT features, which are derived from dense grids instead of difference of Gaussians (DoG) space [50]. Similarly, a 2-D medical image is described using a BoW of SIFT features derived from superpixels [51]. Similar to [44], Zhang *et al.* [52], [53] employed hashing-based retrieval techniques for diagnosis of histopathological images. Two features, SIFT and histograms of oriented gradients [54], are fused and compressed using composite AGH (CAGH) to achieve fast and scalable retrieval. Retrieved database images are then utilized to classify the query image through online learning. Some other CBIR-based CAD methods are investigated in [17] and [55].

III. Proposed Approach

In this section, we first introduce our mammographic ROI retrieval framework based on vocabulary tree, then present the refinement on the weights of tree nodes, and finally, describe how to make a diagnostic decision using the retrieval set. The overview of our approach is shown in Fig. 1.

A. Mammogram Retrieval With a Vocabulary Tree

Our approach builds upon a popular CBIR framework that indexes local image features using vocabulary tree and inverted files [24], [29], [32]. The local feature we choose here is SIFT [26]. 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 [27]. Naturally, the SIFT family are widely adopted by numerous general image

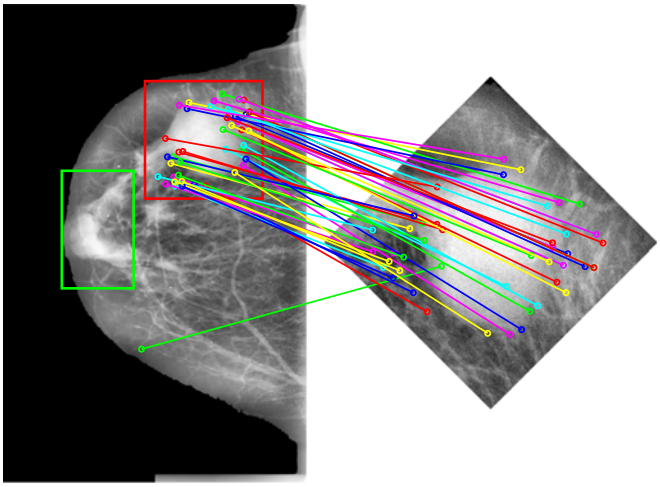

Fig. 2. Matching SIFT features using exhaustive search. The mammogram shown on the left contains two malignant masses within red and green bounding boxes. The mass shown on the right is transformed from the red bounding box using 45° rotation, two-time scale up, 30% contrast enhancement, and addition of 5% Gaussian noise. The SIFT features extracted from the transformed mass are matched with all the features extracted from the mammogram. The 50 matches with largest uniqueness are shown, others are omitted for clarity. 49 of the 50 matches are correct.

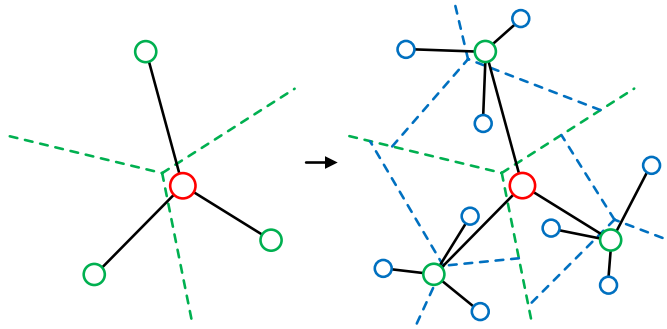

Fig. 3. Training process of a vocabulary tree with depth $L=2$ and branch factor $k=3$. The circles and dashed lines represent the centers and Voronoi boundaries of the clusters, respectively. A cluster center is referred to as a "visual word," and all the visual words form a "vocabulary tree."

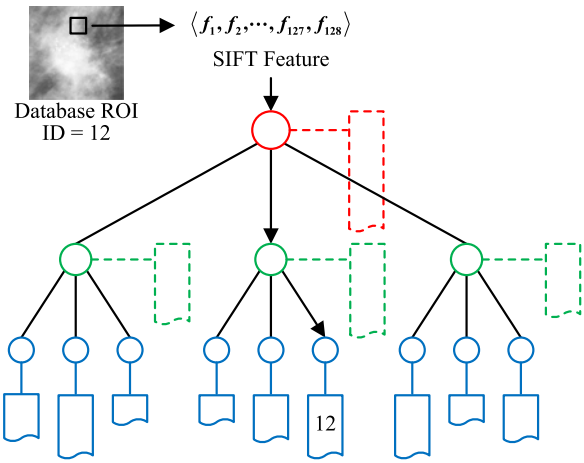

Fig. 4. Quantization and indexing of a database SIFT feature using vocabulary tree and inverted files. Query SIFT features are quantized in the same way, but are not indexed. Each leaf node in the tree has an inverted file (shown in solid lines), which records the IDs of database ROIs containing an instance of the node. An inner node has a virtual inverted file (shown in dashed lines), which is calculated as the concatenation of the files associated with its descendant leaf nodes.

retrieval methods, such as [24], [29] and [32]. They have also been successfully applied to medical image retrieval and analysis [50], [51], [56].

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. An example is given in Fig. 2, which also demonstrates the remarkable robustness and discriminability of SIFT features.

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. 3. 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."

Then, all SIFT features extracted from database ROIs are quantized and indexed using this vocabulary tree and inverted files. As shown in Fig. 4, 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 large-scale 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.

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 [57] 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 speaking, q is represented by a set of paths (features) $q = \{P_i^q\}_{i=1}^m$, where m is the number of features. Each path consists of L nodes $P_i^q = \{v_{i,\ell}^q\}_{\ell=1}^L$, where $v_{i,\ell}^q$ denotes the node on the ℓth level. Similarly, d is represented by $d = \{P_j^d\}_{j=1}^n$, where n is the number of features, and $P_j^d = \{v_{j,\ell}^d\}_{\ell=1}^L$, where $v_{j,\ell}^d$ 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

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

where the normalization factor $1/(m \cdot n)$ 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

$$s_P\left(P_i^q, P_j^d\right) = \sum_{\ell} w\left(v_{i,\ell}^q\right) \cdot \delta\left(v_{i,\ell}^q, v_{j,\ell}^d\right) \tag{2}$$

where w is a weighting function, and δ is the Kronecker delta function, i.e., $\delta(a,b) = 1$ if $a = b$ and $\delta(a,b) = 0$, otherwise. In [29], w is defined following the IDF principle as follows:

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

where N is the total number of database ROIs and N_v 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.

B. 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. As shown in Fig. 5, a majority of features are extracted from normal tissue around a mass. 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 overcounting, inspired by descriptor contextual weighting [24], we incorporate the mammogram-specific node frequencies into IDF scheme to down-weight these features.

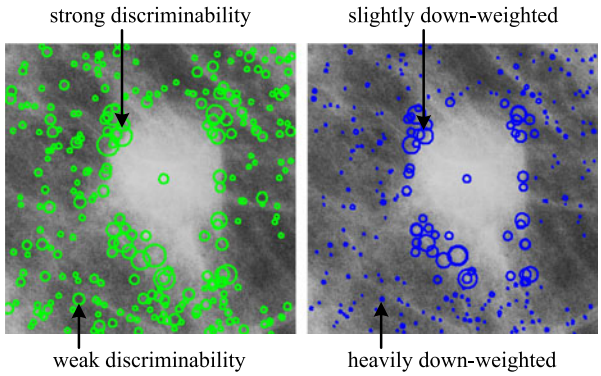

Fig. 5. Effect of adaptive weighting. The left image shows the original IDF weights of the features (only 300 are drawn), and the right image shows the refined weights. The radius of a circle associated with a feature is proportional to its weight.

Suppose the node paths P_i^q of query ROI q and P_j^d of database ROI d have the same node $v \in P_i^q \cap P_j^d = \{v_{i,\ell}^q\}_{\ell=1}^L \cap \{v_{j,\ell}^d\}_{\ell=1}^L$, the node's weight $w(v)$ in (3) is modified to

$$w_{i,j}^{q,d}(v) = w_P\left(P_i^q\right) \cdot w_P\left(P_j^d\right) \cdot idf(v) \tag{4}$$

where the adaptive weight factors $w_P\left(P_i^q\right)$ and $w_P\left(P_j^d\right)$ are calculated based on the frequencies of nodes along paths P_i^q and P_j^d, respectively. Specifically, let $tf\left(v_{i,\ell}^q, q\right)$ be the TF of $v_{i,\ell}^q$ in q, i.e., the number of paths of q that pass through node $v_{i,\ell}^q$, $w_P\left(P_i^q\right)$ is defined as

$$w_P\left(P_i^q\right) = \sqrt{\frac{\sum_\ell w\left(v_{i,\ell}^q\right)}{\sum_\ell w\left(v_{i,\ell}^q\right) \cdot tf\left(v_{i,\ell}^q, q\right)}} \tag{5}$$

where $w(v_{i,\ell}^q)$ is a weighting coefficient, usually set to $idf(v_{i,\ell}^q)$ empirically. $w_P(P_j^d)$ is defined in the same way. The square root in the aforementioned definition is due to the weighting of both $w_P(P_i^q)$ and $w_P(P_j^d)$.

Note that $w_P\left(P_i^q\right)$ is shared for all nodes $v_{i,\ell}^q$ along path P_i^q. In order to determine the importance of a feature P_i^q, $w_P\left(P_i^q\right)$ takes into account the features in q quantized to neighbor tree leaves since they also contribute to $tf(v_{i,\ell}^q, q)$. Consequently, nodes in a subtree with more features are heavily down weighted. The effect of adaptive weighting is illustrated in Fig. 5.

C. Diagnosis of Mammographic Masses

After the retrieval stage, a query mammographic ROI is classified according to its best matched database ROIs using majority logic. Currently, our aim is to distinguish between mass and normal tissue. Malignant and benign masses are not discriminated, since they could be visually indistinguishable and need to be diagnosed through other methods such as biopsy.

Formally speaking, let $\{d_i\}_{i=1}^K$ denote the top K similar database ROIs for q, each d_i has a class tag $c(d_i) \in \{\oplus, \ominus\}$, with the label $\oplus$ for mass and $\ominus$ for normal tissue. q is classified by a weighted majority vote of $\{d_i\}_{i=1}^K$, where the contribution

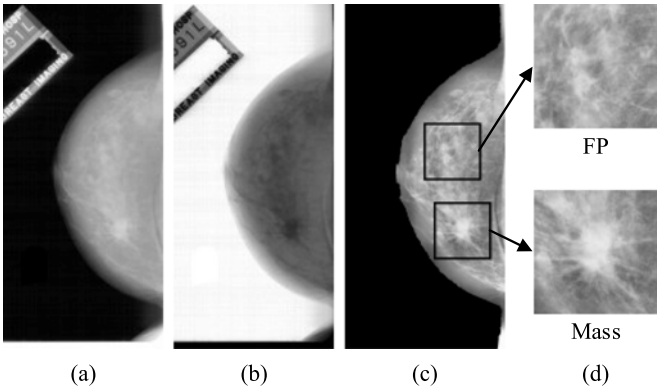

Fig. 6. Construction of our dataset. (a) Original mammogram in gray level format. (b) Normalized mammogram in optical density format. (c) Visually enhanced mammogram. (d) Radiologist-annotated mass and CAD-generated false positive.

of d_i is weighted by its similarity to q

$$c(q) = \arg\max_{c} \sum_{i} s(q, d_i) \cdot \delta(c, c(d_i)). \quad (6)$$

Note that our strategy is similar to the weighted k-nearest neighbor (k-NN) classifier used in [39] and [41].

IV. Experiments

This section validates the proposed mammogram retrieval and diagnosis approach. First, experimental settings, including dataset, compared methods, and evaluation environment, are described. Then, experimental results are presented and analyzed. Finally, impact of parameters is discussed.

A. Experimental Settings

Our experimental dataset is constructed from the digital database for screening mammography (DDSM) [58], [59]. 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.

To simulate practical scenario, a series of ROIs depicting masses and suspicious normal tissues are extracted following the conventions in [21], [36] and [39]. This process is demonstrated in Fig. 6. First of all, mammograms are mapped from gray level to optical density according to DDSM's instructions [59] to eliminate visual difference caused by different scanners [49], [60]. Second, normalized mammograms are processed for better visual quality using inversion, breast segmentation, and contrast enhancement. Third, 2 340 ROIs centered on masses are extracted. Fourth, 9 213 false positives asserted by a CAD system from healthy cases are used as normal regions. This CAD system is based on a cascade of boosted Haar classifiers [61] and trained on a separate mammogram dataset. Note that compared with experiments, which randomly select normal regions [35], our experiment setting is more consistent with practice and more challenging. Finally, of the aforementioned ROIs, 500 mass ROIs, and 500 normal ROIs are randomly selected as queries. The remaining 1 840 mass ROIs and 8 713 normal ROIs, 10 553 ROIs in total, form a large database. The query and database ROIs are selected from different cases in order to avoid positive bias. For a more reliable performance evaluation, the random selection of query ROIs is repeated for five times. After each selection, all the methods are tested, and the average performance from five runs is reported.

TABLE I
Retrieval Precision at Different K

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

We also implement two other medical image retrieval systems for comparison. The first one, presented in [35] and [36], performs a template matching between query ROI and each database ROI based on normalized mutual information (NMI). Experiments in [36] show that NMI obtains good retrieval precision and best diagnosis accuracy among eight information-theoretic similarity measures. The second one, similar to [50] and [51], represents each ROI with a SIFT BoW and measures the χ^2 distance between query ROI and each database ROI. While SIFT feature is derived from dense grids or superpixels in [50] and [51], our implementation employs the traditional SIFT derived from DoG extrema. This is to better test the vocabulary tree framework under the condition that the same feature is utilized. For this method, a vocabulary containing $k_{\mathrm{BoW}} = 1\,000$ visual words is constructed using k-means clustering. Our method is tested twice, with the adaptive weighting scheme deactivated for the first time, and activated for the second time. Both of them employ a vocabulary tree of branch factor $k = 10$ and depth $L = 6$. These four approaches are denoted as NMI, BoW, VocTree, and VocTree+AdaptWeight in the following analysis.

All the methods are implemented in C++ and evaluated on a high-performance laptop with Intel Core i7 processor (6M cache, 2.40 GHz), 16GB memory, and Windows 7 operating system.

B. Results and Analysis

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 I. Two retrieval sets returned by VocTree+AdaptWeight are provided

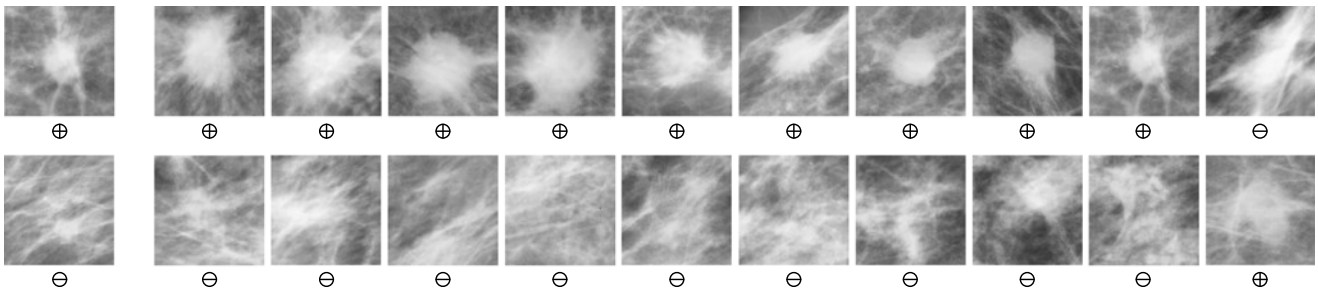

Fig. 7. Two query ROIs (left) and their top $K=10$ retrieved database ROIs calculated by VocTree+AdaptWeight (right). For each ROI, its class is shown below. Both query ROIs are correctly classified according to a weighted majority vote of their retrieval sets.

TABLE II
CLASSIFICATION ACCURACY AT DIFFERENT K

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

TABLE III
PERFORMANCE AT DIFFERENT DATABASE SIZES

Size	Method	Retri. Prec.	Class. Accu.	Time (sec)
2,600	NMI	66.4%	68.8%	3.19
	BoW	70.6%	72.7%	0.71
	VocTree	75.7%	80.1%	**0.29**
	VocTree+AdaptWeight	**81.6%**	**84.4%**	0.34
5,200	NMI	67.9%	69.6%	6.23
	BoW	73.3%	74.5%	1.14
	VocTree	79.6%	81.2%	**0.32**
	VocTree+AdaptWeight	**83.8%**	**86.1%**	0.39
10,553	NMI	70.2%	72.9%	12.31
	BoW	75.5%	76.6%	1.95
	VocTree	81.1%	83.0%	**0.39**
	VocTree+AdaptWeight	**85.4%**	**86.9%**	0.48

in Fig. 7 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.

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 II. 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, VocTree+AdaptWeight achieves a classification accuracy as high as 90.8% at $K = 5$, which is pretty satisfactory.

Finally, *efficiency* and *scalability* are investigated. Efficiency is assessed using the average processing time needed to retrieve and classify a query ROI. Since the classification step is merely a vote on the retrieval set, processing time is actually equal to retrieval time. Besides, as K increases from 1 to 20, a retrieval procedure only needs to change the size of the max/min heap, which records the similarity/distance scores of the retrieved database ROIs. Therefore, the processing time barely changes as K varies, and we only report the time at $K = 20$. Scalability of a method is measured by testing how its performance changes as the database expands. To this end, two smaller databases are constructed by randomly sampling a half and a quarter of database ROIs, and all the methods are evaluated again on these two databases. Their retrieval precisions, classification accuracies, and average processing time at top $K = 20$ retrievals are summarized in Table III. According to this table, we can reach several conclusions. First of all, our methods are consistently superior to the compared approaches with respect to all three evaluation metrics, especially efficiency. VocTree+AdaptWeight obtains even better retrieval precision and classification accuracy than those of VocTree at the cost of a little more processing time. Second, the vocabulary tree framework demonstrates excellent scalability. In this framework, as we explained in Section III-A, the similarity calculation only needs to consider those database features that fall into neighbor leaf nodes as the query features do, which account for a small portion of all the database features. What is more, as the database expands, we can use a larger vocabulary tree (with bigger branch factor k and/or depth L) to reduce the portion of database features that need to be considered. Therefore, the time cost of similarity computation is not only small but also sublinear in database size. On the contrary, NMI and BoW calculate a similarity/distance score between a query ROI and each database ROI, which takes a linear time regarding database size. (The time for query feature extraction and quantization in BoW remains unchanged for different database sizes.) Last but not least, as the database grows, all the methods obtain

TABLE IV
PERFORMANCE OF VOCTREE+ADAPTWEIGHT AT DIFFERENT L

L	Retri. Prec.	Class. Accu.	Time (sec)
3	73.8%	75.3%	1.69
4	78.7%	80.8%	0.91
5	83.1%	85.2%	0.62
6	85.4%	86.9%	0.48
7	**86.2%**	**87.1%**	**0.47**

better retrieval precisions and classification accuracies. This result agrees with the experiments in [21] and [22] and confirms our assumption that it is more likely to find relevant cases and make a correct diagnosis using a large database.

All the experiments lead to several conclusions. 1) NMI obtains the worst results among all the tested methods. The reason is that masses have diverse shapes, sizes, and cluttered background, therefore, it is not suitable to match two entire ROIs without extracting certain features from invariant keypoints. 2) Our method is superior to BoW. Although both employing SIFT feature, they implement different quantization, indexing, and similarity calculation schemes. Specifically, first, BoW uses a single-level k-means clustering for feature quantization, whose computational cost is linear in vocabulary size. As a result, the vocabulary typically has a small size (100 in [50], 1 000 in [51] and our implementation). Instead, our method utilizes hierarchical k-means, whose computational cost is logarithmic in vocabulary size. Thus, it can afford a much larger and more discriminative vocabulary (10^6 in aforementioned experiments). Second, BoW performs exhaustive search without the aid of any index. On the contrary, in our model, quantized database features are indexed using inverted files so that only a small portion of them is considered during similarity computation, and the portion of involved features can be further decreased by increasing vocabulary size (L or k). Actually, experiments on general CBIR datasets demonstrate that our method could retrieve in real time from millions of images. Finally, all the visual words in BoW are treated equally, whereas their weights are elaborately adjusted according to the whole database (IDF) and each query (TF and adaptive weighting). 3) Adaptive weighting, which down weights the excessive features extracted from normal regions, could improve retrieval precision and classification accuracy without considerably reducing efficiency.

C. Discussion of Parameters

To test the impact of parameters on our method's performance, we have trained several vocabulary trees of branch factor $k = 10$ and depth $L = 3, \ldots, 7$. For each vocabulary tree, the performance of VocTree+AdaptWeight at top $K = 20$ retrievals is measured using five randomly selected query sets, and the average performance from five runs is summarized in Table IV. From this table, we can see that the retrieval precision, classification accuracy, and computational efficiency of VocTree+AdaptWeight improve substantially as L goes from 3 to 5, then improve slightly as L increases to 6 and 7. Two conclusions can be drawn from this observation. On the one hand, larger vocabulary trees tend to achieve better performance. As explained in Section III-A, a larger vocabulary tree has smaller and more discriminative leaf nodes, which result in better retrieval precision as well as classification accuracy. Besides, as the total number of leaf nodes increases, the portion of database features that need to be considered during similarity calculation is reduced, therefore, the efficiency is also improved. On the other hand, the vocabulary tree framework could benefit from more training features. The performance gain from $L = 6$ to 7 is very small. It is probably due to the limited number of training features, since nearly a third of leaf nodes are empty during the training process when L becomes 7. These two conclusions are consistent with the observations in general image retrieval [24], [29].

V. CONCLUSION

Mammography has played a key role in the early diagnosis of breast cancer. To facilitate mammographic masses detection, numerous CAD methods are developed, and a growing number of them begin to utilize CBIR techniques. Compared with classifier-based approaches, CBIR-based methods can detect masses of uncommon appearance or size, bypass the obscure mass boundary problem, provide more clinical evidence, and improve PACS systems. However, lack of scalability remains a major drawback of current CBIR-based CAD methods and sets a limit on their retrieval precision as well as diagnostic accuracy.

In this paper, we propose to use scalable CBIR 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.

Future endeavors will be devoted to improve retrieval precision. One possible solution is to utilize several visual features. Specifically, intensity, texture, and shape features can complement the adopted SIFT feature. Most of them are global features and can be indexed using hash tables [30] to achieve sublinear similarity search. In order to combine multiple features, existing methods either concatenate them to form a new feature [37], [39], [41], [42], [48], [49], or aggregate individual retrieval sets according to similarity/distance scores [38] or ranks [47]. However, these approaches use fixed or user-defined parameters, e.g., weight of each feature in similarity calculation. Consequently, they cannot completely integrate the strengths of complementary features, which may work well for different kinds of queries. To overcome this problem, we can employ feature fusion strategy [32], [33], such as graph fusion [32] that adaptively merges

individual retrieval sets through a link analysis on a fused graph. Besides, the proposed method can be applied to other medical domains, such as retrieval and diagnosis of nodules in lung CT images.

REFERENCES

[1] American Cancer Society, *Breast Cancer Facts & Figures 2013-2014*. Atlanta, GA, USA: American Cancer Society, 2013.

[2] N. Howlader, A. M. Noone, M. Krapcho, J. Garshell, N. Neyman, S. F. Altekruse, C. L. Kosary, M. Yu, J. Ruhl, Z. Tatalovich, H. Cho, A. Mariotto, D. R. Lewis, H. S. Chen, E. J. Feuer, and K. A. Cronin, *SEER Cancer Statistics Review, 1975-2010*. National Cancer Institute, Bethesda, MD, USA, 2013.

[3] H.-D. Cheng, X.-J. Shi, R. Min, L.-M. Hu, X.-P. Cai, and H.-N. Du, "Approaches for automated detection and classification of masses in mammograms," *Pattern Recog.*, vol. 39, no. 4, pp. 646–668, 2006.

[4] A. Oliver, J. Freixenet, J. Martí, E. Pérez, 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.

[5] P. Skaane, K. Engedal, and A. Skjennald, "Interobserver variation in the interpretation of breast imaging," *Acta Radiol.*, vol. 38, no. 4, pp. 497–502, 1997.

[6] R. M. Rangayyan, F. J. Ayres, and J. E. Leo Desautels, "A review of computer-aided diagnosis of breast cancer: Toward the detection of subtle signs," *J. Franklin Inst.*, vol. 344, pp. 312–348, 2007.

[7] R. L. Birdwell, D. M. Ikeda, K. F. O'Shaughnessy, and E. A. Sickles, "Mammographic characteristics of 115 missed cancers later detected with screening mammography and the potential utility of computer-aided detection," *Radiology*, vol. 219, no. 1, pp. 192–202, 2001.

[8] F. Winsberg, M. Elkin, J. Macy, V. Bordaz, and W. Weymouth, "Detection of radiographic abnormalities in mammograms by means of optical scanning and computer analysis," *Radiology*, vol. 89, no. 2, pp. 211–215, 1967.

[9] J. Tang, R. M. Rangayyan, J. Xu, I. El Naqa, and Y. Yang, "Computer-aided detection and diagnosis of breast cancer with mammography: Recent advances," *IEEE Trans. Inf. Technol. Biomed.*, vol. 13, no. 2, pp. 236–251, Mar. 2009.

[10] 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.

[11] 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.

[12] Y. Rui, T. S. Huang, and S.-F. Chang, "Image retrieval: Current techniques, promising directions, and open issues," *J. Vis. Commun. Image Represent.*, vol. 10, no. 1, pp. 39–62, 1999.

[13] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 12, pp. 1349–1380, Dec. 2000.

[14] F. Long, H. Zhang, and D. D. Feng, "Fundamentals of content-based image retrieval," in *Multimedia Information Retrieval and Management*. New York, NY, USA: Springer, 2003, pp. 1–26.

[15] Y. Liu, D. Zhang, G. Lu, and W.-Y. Ma, "A survey of content-based image retrieval with high-level semantics," *Pattern Recog.*, vol. 40, no. 1, pp. 262–282, 2007.

[16] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Image retrieval: Ideas, influences, and trends of the new age," *ACM Comput. Surveys*, vol. 40, no. 2, 2008.

[17] H. Müller, N. Michoux, D. Bandon, and A. Geissbühler, "A review of content-based image retrieval systems in medical applications - clinical benefits and future directions," *Int. J. Med. Informat.*, vol. 73, no. 1, pp. 1–23, 2004.

[18] I. El-Naqa, Y. Yang, N. P. Galatsanos, R. M. Nishikawa, and M. N. Wernick, "A similarity learning approach to content-based image retrieval: Application to digital mammography," *IEEE Trans. Med. Imag.*, vol. 23, no. 10, pp. 1233–1244, Oct. 2004.

[19] C. J. D'Orsi, E. A. Sickles, E. B. Mendelson, E. A. Morris *et al.*, *ACR BI-RADS Atlas, Breast Imaging Reporting and Data System*. 5th ed., Reston, VA, USA: Amer. College Radiol., 2013.

[20] J. Wood, T. Andersson, A. Bachem, C. Best, F. Genova, D. R. Lopez, W. Los, M. Marinucci, L. Romary, H. Van de Sompel, J. Vigen, and P. Wittenburg, *Riding the Wave: How Europe can Gain from the Rising Tide of Scientific Data*, European Union, Brussels, Belgium, 2010.

[21] S. C. Park, R. Sukthankar, L. Mummert, M. Satyanarayanan, and B. Zheng, "Optimization of reference library used in content-based medical image retrieval scheme," *Med. Phys.*, vol. 34, no. 11, pp. 4331–4339, 2007.

[22] G. D. Tourassi, B. Harrawood, S. Singh, and J. Y. Lo, "Information-theoretic CAD system in mammography: Entropy-based indexing for computational efficiency and robust performance," *Med. Phys.*, vol. 34, no. 8, pp. 3193–3204, 2007.

[23] G. Langs, A. Hanbury, B. H. Menze, and H. Müller, "VISCERAL: Towards large data in medical imaging—Challenges and directions," in *Proc. Med. Content-Based Retrieval Clin. Decision Support*, 2012, pp. 92–98.

[24] X. Wang, M. Yang, T. Cour, S. Zhu, K. Yu, and T. X. Han, "Contextual weighting for vocabulary tree based image retrieval," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 209–216.

[25] M. Jiang, S. Zhang, J. Liu, T. Shen, and D. N. Metaxas, "Computer-aided diagnosis of mammographic masses using vocabulary tree-based image retrieval," in *Proc. IEEE Int. Symp. Biomed. Imag.*, 2014, pp. 1123–1126.

[26] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[27] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 10, pp. 1615–1630, Oct. 2005.

[28] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2003, pp. 1470–1477.

[29] D. Nistér and H. Stewénius, "Scalable recognition with a vocabulary tree," in *Proc. IEEE Comput. Vision Pattern Recog.*, 2006, pp. 2161–2168.

[30] A. Gionis, P. Indyk, and R. Motwani, "Similarity search in high dimensions via hashing," in *Proc. Int. Conf. Very Large Data Bases*, 1999, pp. 518–529.

[31] W. Liu, J. Wang, S. Kumar, and S.-F. Chang, "Hashing with graphs," in *Proc. Int. Conf. Mach. Learn.*, 2011, pp. 1–8.

[32] S. Zhang, M. Yang, T. Cour, K. Yu, and D. N. Metaxas, "Query specific fusion for image retrieval," in *Proc. Eur. Conf. Comput. Vision*, 2012, pp. 660–673.

[33] S. Zhang, J. Huang, H. Li, and D. N. Metaxas, "Automatic image annotation and retrieval using group sparsity," *IEEE Trans. Syst. Man, Cybern. B, Cybern.*, vol. 42, no. 3, pp. 838–849, Jun. 2012.

[34] Y. Rui, T. S. Huang, M. Ortega, and S. Mehrotra, "Relevance feedback: A power tool for interactive content-based image retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 8, no. 5, pp. 644–655, Sep. 1998.

[35] G. D. Tourassi, R. Vargas-Voracek, D. M. Catarious Jr, and C. E. Floyd Jr, "Computer-assisted detection of mammographic masses: A template matching scheme based on mutual information," *Med. Phys.*, vol. 30, no. 8, pp. 2123–2130, 2003.

[36] G. D. Tourassi, B. Harrawood, S. Singh, J. Y. Lo, and C. E. Floyd, "Evaluation of information-theoretic similarity measures for content-based retrieval and detection of masses in mammograms," *Med. Phys.*, vol. 34, no. 1, pp. 140–150, 2007.

[37] H. Alto, R. M. Rangayyan, and J. E. L. Desautels, "Content-based retrieval and analysis of mammographic masses," *J. Electron. Imag.*, vol. 14, no. 2, pp. 023 016–1–023 016–17, 2005.

[38] Y. Tao, S.-C. B. Lo, M. T. Freedman, and J. Xuan, "A preliminary study of content-based mammographic masses retrieval," *Proc. SPIE*, vol. 6514, pp. 1–12, 2007.

[39] B. Zheng, A. Lu, L. A. Hardesty, J. H. Sumkin, C. M. Hakim, M. A. Ganott, and D. Gur, "A method to improve visual similarity of breast masses for an interactive computer-aided diagnosis environment," *Med. Phys.*, vol. 33, no. 1, pp. 111–117, 2006.

[40] C.-H. Wei, Y. Li, and P. J. Huang, "Mammogram retrieval through machine learning within BI-RADS standards," *J. Biomed. Inform.*, vol. 44, no. 4, pp. 607–614, 2011.

[41] F. Narváez, G. Díaz, and E. Romero, "Multi-view information fusion for automatic BI-RADS description of mammographic masses," *Proc. SPIE*, vol. 7963, pp. 79 630A-1–79 630A-7, 2011.

[42] F. Narváez, G. Díaz, F. Gómez, and E. Romero, "A content-based retrieval of mammographic masses using a curvelet descriptor," *Proc. SPIE*, vol. 8315, pp. 83 150A-1–83 150A-7, 2012.

[43] S. K. Kinoshita, P. M. de Azevedo-Marques, R. R. Pereira Jr, J. A. H. Rodrigues, and R. M. Rangayyan, "Content-based retrieval of mammograms using visual features related to breast density patterns," *J. Digit. Imag.*, vol. 20, no. 2, pp. 172–190, 2007.

[44] 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.
[45] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vis.*, vol. 42, no. 3, pp. 145–175, 2001.
[46] W. Cai, D. D. Feng, and R. R. Fulton, "Content-based retrieval of dynamic PET functional images," *IEEE Trans. Inf. Technol. Biomed.*, vol. 4, no. 2, pp. 152–158, Jun. 2000.
[47] J. Kim, W. Cai, D. D. Feng, and H. Wu, "A new way for multidimensional medical data management: Volume of interest (VOI)-based retrieval of medical images with visual and functional features," *IEEE Trans. Inf. Technol. Biomed.*, vol. 10, no. 3, pp. 598–607, Jul. 2006.
[48] A. M. Aisen, L. S. Broderick, H. Winer-Muram, C. E. Brodley, A. C. Kak, C. Pavlopoulou, J. Dy, C.-R. Shyu, and A. Marchiori, "Automated storage and retrieval of thin-section CT images to assist diagnosis: System description and preliminary assessment," *Radiology*, vol. 228, no. 1, pp. 265–270, 2003.
[49] M. E. Mattie, L. H. Staib, E. Stratmann, H. D. Tagare, J. S. Duncan, and P. L. Miller, "Pathmaster: Content-based cell image retrieval using automated feature extraction," *J. Amer. Med. Inform. Assoc.*, vol. 7, no. 4, pp. 404–415, 2000.
[50] B. André, T. Vercauteren, A. M. Buchner, M. B. Wallace, and N. Ayache, "Learning semantic and visual similarity for endomicroscopy video retrieval," *IEEE Trans. Med. Imag.*, vol. 31, no. 6, pp. 1276–1288, Jun. 2012.
[51] S. Haas, R. Donner, A. Burner, M. Holzer, and G. Langs, "Superpixel-based interest points for effective bags of visual words medical image retrieval," in *Proc. Med. Content-Based Retrieval Clin. Decision Support*, 2011, pp. 58–68.
[52] X. Zhang, W. Liu, M. Dundar, S. Badve, and S. Zhang, "Towards large-scale histopathological image analysis: Hashing-based image retrieval," *IEEE Trans. Med. Imag.*, to be published.
[53] X. Zhang, L. Yang, W. Liu, H. Su, and S. Zhang, "Mining histopathological images via composite hashing and online learning," in *Proc. Med. Image Comput. Comput.-Assisted Intervention Conf.*, 2014, pp. 479–486.
[54] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Vision Pattern Recog.*, 2005, pp. 886–893.
[55] A. Kumar, J. Kim, W. Cai, M. Fulham, and D. Feng, "Content-based medical image retrieval: A survey of applications to multidimensional and multimodality data," *J. Digit. Imag.*, vol. 26, no. 6, pp. 1025–1039, 2013.
[56] J. C. Caicedo, A. Cruz, and F. A. Gonzalez, "Histopathology image classification using bag of features and kernel functions," in *Proc. Conf. Artif. Intell. Med.*, 2009, pp. 126–135.
[57] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Inf. Process. Manag.*, vol. 24, no. 5, pp. 513–523, 1988.
[58] M. Heath, K. Bowyer, D. Kopans, W. P. Kegelmeyer Jr, R. Moore, K. Chang, and S. Munishkumaran, "Current status of the digital database for screening mammography," in *Digital Mammography*. Dordrecht, Netherlands, Springer, 1998, pp. 457–460.
[59] University of South Florida,. (2001). USF digital mammography home page. [Online]. Available: http://marathon.csee.usf.edu/Mammography/Database.html
[60] Y. Wang, B. M. Keller, Y. Zheng, R. J. Acciavatti, J. C. Gee, A. D. A. Maidment, and D. Kontos, "A phantom study for assessing the effect of different digital detectors on mammographic texture features," in *Proc. Int. Workshop Digital mammography*, 2012, pp. 604–610.
[61] P. A. Viola and M. J. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Comput. Vision Pattern Recog.*, 2001, pp. I-511–I-518.

Authors' photographs and biographies not available at the time of publication.