Towards Large-Scale Histopathological Image Analysis: Hashing-Based Image Retrieval

Xiaofan Zhang, *Student Member, IEEE*, Wei Liu, *Member, IEEE*, Murat Dundar, *Member, IEEE*, Sunil Badve, and Shaoting Zhang*, *Member, IEEE*

***Abstract*—Automatic analysis of histopathological images has been widely utilized leveraging computational image-processing methods and modern machine learning techniques. Both computer-aided diagnosis (CAD) and content-based image-retrieval (CBIR) systems have been successfully developed for diagnosis, disease detection, and decision support in this area. Recently, with the ever-increasing amount of annotated medical data, large-scale and data-driven methods have emerged to offer a promise of bridging the semantic gap between images and diagnostic information. In this paper, we focus on developing *scalable* image-retrieval techniques to cope intelligently with massive histopathological images. Specifically, we present a supervised kernel hashing technique which leverages a small amount of supervised information in learning to compress a 10 000-dimensional image feature vector into only tens of binary bits with the informative signatures preserved. These binary codes are then indexed into a hash table that enables real-time retrieval of images in a large database. Critically, the supervised information is employed to bridge the semantic gap between low-level image features and high-level diagnostic information. We build a scalable image-retrieval framework based on the supervised hashing technique and validate its performance on several thousand histopathological images acquired from breast microscopic tissues. Extensive evaluations are carried out in terms of image classification (i.e., benign versus actionable categorization) and retrieval tests. Our framework achieves about 88.1% classification accuracy as well as promising time efficiency. For example, the framework can execute around 800 queries in only 0.01 s, comparing favorably with other commonly used dimensionality reduction and feature selection methods.**

***Index Terms*—Breast lesion, hashing, high dimension, histopathological image analysis, large-scale image retrieval, supervised learning.**

I. Introduction

BREAST cancer is the second most common cancer in women. In the United States, breast cancer alone accounted for 29% of all new cancer cases among women in 2013. Fortunately, over the past decade the death rate for female breast cancer has decreased by 17%, although the incidence of breast cancer has risen. This situation is largely the result of improvements in early detection and treatment [1]. The preinvasive stages of breast cancer have two general histological categories: the lobular and ductal subtypes. Most work focuses on ductal preinvasive cancer, since approximately 80% of all diagnosed preinvasive and invasive breast cancers are of this subtype [2]. Pioneering work done by Page & Dupont divided intraductal lesions into three major classes [usual ductal hyperplasia (UDH), atypical ductal hyperplasia (ADH), and ductal carcinoma *in situ* (DCIS)] and suggested that ADH and DCIS should be considered as precursor lesions [3]. On the other hand, there is little evidence to suggest that UDH is a precursor lesion. These differences in biology are manifested in the treatment of patients. The standard of care for patients diagnosed with UDH on core biopsy is routine follow-up, whereas those with ADH and DCIS are subjected to excisional biopsy, which can be associated with pain, discomfort, and scarring for the patient while also adding significantly to health-care costs.

Classical examination methods include screening tests and biopsy. Compared to mammography, histopathology slides provide more comprehensive information for diagnosis. However, manual examination of microscopic images is labor intensive and time consuming, and may depend on a subjective assessment by the pathologist, which poses a special challenge in the diagnosis of preinvasive breast cancer.

Computer-aided diagnosis (CAD) systems have been widely used in an attempt to relieve the workload on pathologists and to offer more reliable and consistent analysis of histopathology images. The appearance and widespread study of CAD systems for cytological diagnosis can be traced back to the last century [4]. Recently some studies have focused on analyzing high-resolution images digitized from tissue histopathology slides [5]–[11].

In addition to the classifier-based CAD systems providing diagnosis results or grading scores, content-based image retrieval (CBIR) has also been extensively investigated for decision support in many clinical applications, including digital pathology [12]–[16]. Given an image database with diagnosis information, CBIR methods aim to retrieve and visualize images with morphological profiles most relevant to and consistent with the query image [17]. CBIR can also be used for classification purposes by considering the majority diagnosis of the retrieved images as the most likely diagnosis.

Traditional CBIR methods in medical files usually focus on small data sets that have only tens or hundreds of images. New opportunities and challenges arise with the ever-in-

Manuscript received August 18, 2014; revised September 24, 2014; accepted September 29, 2014. Date of publication October 09, 2014; date of current version January 30, 2015. This work was supported in part by the University of North Carolina at Charlotte, Charlotte Research Institute (CRI) , and in part by the Oak Ridge Associated Universities for the Ralph E. Powe Junior Faculty Enhancement Award . *Asterisk indicates corresponding author.*

X. Zhang is with the Department of Computer Science, University of North Carolina, Charlotte, NC 28027 USA (e-mail: xzhang35@uncc.edu).

W. Liu is with the IBM T. J. Watson Research Center, Yorktown Heights, NY 10598 USA (e-mail: weiliu@us.ibm.com).

M. Dundar is with the Computer and Information Science Department, Indiana University–Purdue University, Indianapolis, IN 46202 USA.

S. Badve is with the Department of Pathology, Indiana University, Indianapolis, IN 46202 USA.

*S. Zhang is with the Department of Computer Science, University of North Carolina, Charlotte, NC 28027 USA (e-mail: szhang16@uncc.edu).

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/TMI.2014.2361481

creasing amount of patient data in the current era. Intuitively, larger databases provide more comprehensive information and may improve the accuracy of CBIR systems. On the other hand, achieving an acceptable retrieval efficiency is a challenging task for large-scale data, especially when very large numbers of features are required to capture subtle image descriptors. In fact, CBIR methods usually suffer from the "curse of dimensionality" and low computational efficiency when using high-dimensional features in large databases. Although cloud and grid computing are a potential solution for efficient computing [18], [19], few efforts have been made to develop computational and scalable algorithms for large-scale histopathological image analysis.

In this paper, we focus on scalable image-retrieval methods for the image-guided diagnosis of preinvasive breast cancer using several thousand breast-tissue images. In particular, we design a CBIR framework by leveraging high-dimensional texture features and hashing-based methods [20]–[24] for large-scale image retrieval. A kernel-based supervised hashing model is introduced to encode a high-dimensional image feature vector to short binary bits using only a limited number of labeled images. Such binary bits can significantly reduce the amount of memory required for storing the image database. The proposed technique also allows for real-time querying of an image collection of millions of images, thanks to a hash table containing only binary codes. A preliminary version of this work was reported in a conference paper [25], with 500 testing images. In this paper, we improve the optimization scheme for efficiency and online updating, and also validate the proposed framework thoroughly on a larger data set. Specifically, we validate our framework in terms of both image classification and retrieval on a breast-lesion data set containing 3121 images from 116 patients and achieve an accuracy of 88.1% in a 10-ms query time for around 800 testing images and a precision of 83% in retrieval.

The major contribution of this work is twofold. 1) A comprehensive and large-scale CBIR framework designed to analyze histopathological images by leveraging high-dimensional texture features and hashing-based methods is successfully employed for the image-guided diagnosis of preinvasive breast cancer using several thousand breast-tissue images. 2) An efficient updating and optimization scheme is proposed to improve the supervised and kernelized hashing method. As a benefit, our framework can handle new training samples more efficiently than the traditional method.

The paper is organized as follows. Section II reviews relevant work in two categories, CAD systems and CBIR systems. Section III introduces the formulation and optimization of our hashing-based image-retrieval framework for decision support. Section IV presents the experimental results. Section V summarizes our contributions and plans for future work.

II. Related Work

In this section, we review three major categories of relevant work: CAD systems for the analysis of histopathological images, CBIR systems in medical image analysis, and hashing methods for large-scale image retrieval.

A. Classifier-Based CAD for Histopathological Images

Classifier-based CAD systems consist mainly of image preprocessing, detection and/or segmentation, feature extraction, machine learning-based classification, and postprocessing methods. We briefly review relevant work mainly according to the classification method, which is the focus of this paper. For example, Petushi *et al.* [26] employed adaptive thresholding and morphological operations to segment cells and represent high-density areas of different types of nuclei. These cells were then classified with linear discriminant analysis (LDA) and forward/backward search methods. Yang *et al.* [27] employed filter banks to model phenotypic appearance in histopathological images, which were classified via a gentle boosting mechanism. Caicedo *et al.* [28] proposed to use SIFT to detect key points and extract local descriptors, which are used to obtain a bag-of-words [29] and classified using a support vector machine (SVM) with kernel functions. Basavanhally *et al.* [30] proposed to detect locations of nuclei using a combination of region growing and Markov random fields. Three graphs (i.e., Voronoi diagram, Delaunay triangulation, and minimum spanning tree) are constructed to describe the arrangement of cells. SVM is then employed to classify the high or low presence of lymphocytic infiltration that can be used to evaluate phenotypic changes in breast cancer. Dundar *et al.* [31] proposed to segment cells using a Gaussian mixture model (GMM) and a watershed algorithm and to describe individual cells by their size, shape, and nucleoli. After that, multiple-instance learning (MIL) with SVM was used to identify and classify the stage of breast lesion. Most of these methods use one type of features to classify histopathological images. It is also possible to fuse multiple features in classifiers for comprehensive information. For instances, Tabesh *et al.* [32] aggregated color, texture, and morphometric cues at the global and histological object levels for k-nearest neighbors (kNN) and SVM-based classification. Doyle *et al.* [33] graded breast cancers with both graph-based and texture features for SVM-based classification.

B. CBIR Systems for Medical Image Analysis

CBIR shows its importance in medical image analysis by providing doctors with diagnostic aid in the form of visualizing existing and relevant cases, along with diagnosis information. Therefore, clinical decision-support techniques such as case-based reasoning or evidence-based medicine have a strong need for retrieving images that can be valuable for diagnosis. For example, Comaniciu *et al.* [12] proposed a content-based image-retrieval system that supports decision making in clinical pathology, in which a central module and fast color segmenter are used to extract features such as shape, area, and texture of the nucleus. System performance was assessed through a ten-fold cross-validated classification and compared with that of a human expert on a database containing 261 digitized specimens. Dy *et al.* [34] described a new hierarchical approach of CBIR based on multiple feature sets and a two-step approach. The query image is classified into different classes with best discriminative features between the classes. Then similar images are searched in the predicted class with the features customized to distinguish subclasses. El-Naqa *et al.* [35] proposed a hierarchical learning approach that consists

of a cascade of a binary classifier and a regression module to optimize retrieval effectiveness and efficiency. They applied this to retrieve digital mammograms and evaluated it on a database of 76 mammograms. Greenspan *et al.* [36] proposed a CBIR system that consists of a continuous and probabilistic image-representation scheme. It uses GMM and information-theoretic image matching via the Kullback-Leibler (KL) measure to match and categorize X-ray images by body region. Song *et al.* [37] designed a hierarchical spatial matching-based image-retrieval method using spatial pyramid matching to effectively extract and represent the spatial context of pathological tissues. In the context of histopathological images from breast tissues, Schnorrenberg *et al.* [38] extended the biopsy analysis support system to include indexing and content-based retrieval of biopsy slide images. A database containing 57 breast-cancer cases was used for evaluation. Zheng *et al.* [13] designed a CBIR system to retrieve images and their associated annotations from a networked microscopic pathology image database based on four types of image features. Akakin *et al.* [15] proposed a CBIR system using the multi-tiered approach to classify and retrieve microscopic images, which enables both multi-image query and slide-level image retrieval in order to protect the semantic consistency among the retrieved images.

As emphasized in [39], scalability is the key factor in CBIR for medical image analysis. However, owing to the difficulties in developing scalable CBIR systems for large-scale data sets, most previous systems have been tested on a relatively small number of cases. With the goal of comparing CBIR methods on a larger scale, ImageCLEF and VISCERAL provide benchmarks for medical image-retrieval tasks [40]–[44]. Recently, Foran *et al.* [19] designed a CBIR system named ImageMiner for comparative analysis of tissue microarrays by harnessing the benefits of high-performance computing and grid technology. However, few attempts have been made to design computational and scalable retrieval algorithms in this area, particularly for the analysis of histopathological images.

C. Hashing Methods for Large-Scale Image Retrieval

Recently, hashing methods have been intensively investigated in the machine learning and computer vision community for large-scale image retrieval. Representative methods include, but are not limited to, weakly-supervised hashing in kernel space [45], semi-supervised hashing [46] supervised hashing [24], and compact kernel hashing with multiple features [47]. Among these methods, kernelized and supervised hashing (KSH) [24] is generally considered the most effective, achieving state-of-the-art performance with a moderate training cost. Therefore, this was chosen in our framework for scalable image retrieval. The central idea of KSH is to reduce the gap between low-level hash code similarity and high-level semantic (label) similarity by virtue of supervised training. In doing so, a similarity search in the binary code space can reveal the given semantics of examples. In other words, KSH does well in incorporating the given semantics into the learned hash functions or codes, while the other hashing methods are inadequate in leveraging the semantics. Specifically, compared to the unsupervised kernel hashing method [22], [24] and the semi-supervised linear hashing method [23], [46], KSH shows much higher search accuracy, as it takes full advantage of supervised information (originating from the semantics) that is not well exploited by those unsupervised and semi-supervised methods. Even compared against competing supervised hashing methods such as binary reconstructive embedding (BRE) [48] and minimal loss hashing (MLH) [49], KSH still shows clear accuracy gains yet with much shorter training time. One major limitation of KSH is that optimization in order to obtain accurate hash functions is very time consuming. Therefore, we propose to further improve the original KSH with an efficient updating and optimization scheme that handles new training samples on the fly.

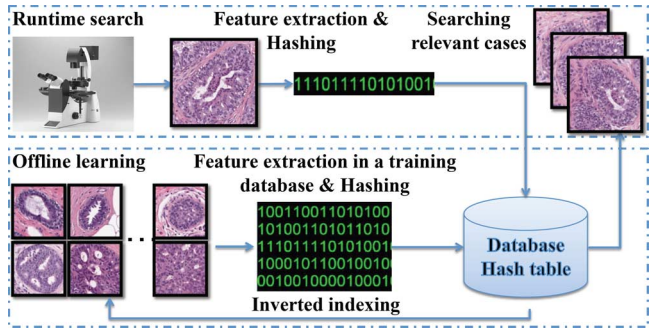

Fig. 1. Overview of our proposed system.

III. Methodology

A. Overview of Scalable Image Retrieval Framework

Fig. 1 shows the proposed framework of our scalable image retrieval-based diagnosis system. It includes offline learning and run-time search. During the offline learning, we first extract high-dimensional visual features from digitized histopathological images. These features model texture and appearance information based on SIFT [50] and are quantized with a bag-of-words [29]. The SIFT descriptor is an effective local texture feature that uses the difference of Gaussian (DoG) detection result and considers the gradient of pixels around the detected region. It can provide an informative description of cell appearance and is robust to subtle changes in staining color. It has been used in both general computer vision tasks and histopathological image analysis [28], [51].

Although these features can be used directly to measure the similarity among images, computational efficiency is an issue, especially when searching in a large database (e.g., exhaustively searching k-nearest neighbors). Therefore, we employ a hashing method to compress these features into binary codes with tens of bits. Such short binary features allow easy mapping into a hash table for real-time search. Each feature is then linked to the corresponding training images using an inverted index. During a run-time query, high-dimensional features are extracted from the query image and then projected to the binary codes. With a hash table, searching for nearest neighbors can be achieved in a constant time, irrespective of the number of images. The retrieved images (via inverted indices of nearest neighbors) can be used to interpret this new case or for decision support based on majority voting.

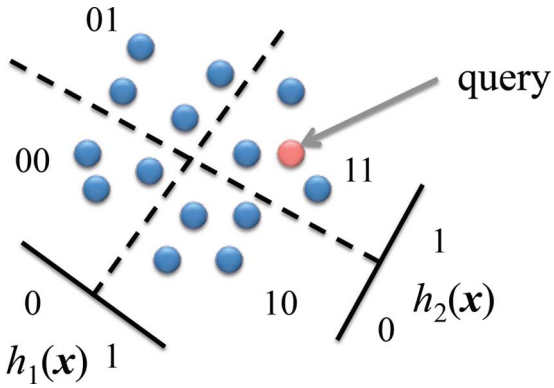

Fig. 2. Visualization of desirable hash functions as a hyperplane.

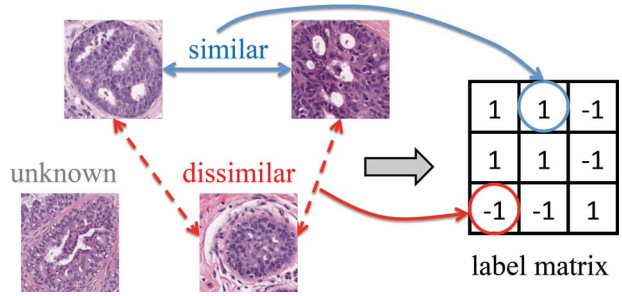

Fig. 3. Supervised information is encoded in the label matrix S.

B. Kernelized and Supervised Hashing

In this section, we introduce the key module for histopathological image retrieval, a kernelized and supervised hashing method.

1) Hashing Method: Given a set of image feature vectors $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\} \subset \mathbb{R}^d$ (in our case, $\boldsymbol{x}_i$ is the high-dimensional texture feature extracted from the ith histopathological image), a hashing method aims to find a group of proper hash functions $h : \mathbb{R}^d \mapsto \{1, -1\}^1$, each of which generates a single hash bit to preserve the similarity of original features. Searching k-nearest neighbors using tens of bits is significantly faster than traditional methods (e.g., Euclidean distance-based brute-force search), owing to constant-time hash-table lookups and/or efficient Hamming distance computation. Note that hashing methods are different from dimensionality-reduction techniques, since a fundamental requirement of hashing is to map similar feature vectors into the same bucket with high probability. Fig. 2 visualizes desirable hash functions as a hyperplane to separate higher-dimensional features. Therefore, hashing methods need to ensure that the generated hash bits have balanced and uncorrelated bit distributions, which leads to maximum information at each single bit and minimum redundancy among all bits.

2) Kernelized Hashing: Kernel methods can handle practical data that are mostly linearly inseparable. For histopathological images, linear inseparability is an important constraint that needs to be taken into account when building hashing methods. Therefore, kernel functions should be considered in hashing methods $h = sgn(f(x))$ [22] to map the feature vectors into higher-dimensional space. A kernel function is denoted as $\kappa : \mathbb{R}^d \times \mathbb{R}^d \mapsto \mathbb{R}$. The prediction function $f : \mathbb{R}^d \mapsto \mathbb{R}$ with kernel κ plugged in is defined as

$$f(\boldsymbol{x}) = \sum_{j=1}^{m} \kappa(\boldsymbol{x}_{(j)}, \boldsymbol{x}) a_j - b \tag{1}$$

where $\boldsymbol{x}_{(1)}, \ldots, \boldsymbol{x}_{(m)}$ are $m (m \ll n)$ feature vectors randomly selected from $\mathcal{X}$, $a_j \in \mathbb{R}$ is the coefficient, and $b \in \mathbb{R}$ is the bias.

The bits generated from hash functions h using f aim to keep as much information as possible, so the hash functions should produce a balanced distribution of bits, i.e., $\sum_{i=1}^{n} h(\boldsymbol{x}_i) = 0$. Therefore, b is set as the median of $\{\sum_{j=1}^{m} \kappa(\boldsymbol{x}_{(j)}, \boldsymbol{x}_i) a_j\}_{i=1}^{n}$, which is usually approximated by the mean. Adding this constraint into (1), we obtain

$$f(\boldsymbol{x}) = \sum_{j=1}^{m} \left(\kappa(\boldsymbol{x}_{(j)}, \boldsymbol{x}) - \frac{1}{n} \sum_{i=1}^{n} \kappa(\boldsymbol{x}_{(j)}, \boldsymbol{x}_i) \right) a_j = \boldsymbol{a}^\top \bar{\boldsymbol{k}}(\boldsymbol{x}) \tag{2}$$

where $\boldsymbol{a} = [a_1, a_2, \ldots, a_m]^\top$. $\bar{\boldsymbol{k}} : \mathbb{R}^d \mapsto \mathbb{R}^m$ is $\bar{\boldsymbol{k}}(\boldsymbol{x}) = \left[\kappa(\boldsymbol{x}_{(1)}, \boldsymbol{x}) - \mu_1, \ldots, \kappa(\boldsymbol{x}_{(m)}, \boldsymbol{x}) - \mu_m\right]^\top$, in which $\mu_j = \sum_{i=1}^{n} \kappa(\boldsymbol{x}_{(j)}, \boldsymbol{x}_i)/n$.

The vector $\boldsymbol{a}$ is the most important factor that determines hash functions. In traditional kernelized hashing methods, $\boldsymbol{a}$ is defined as a random direction drawn from a Gaussian distribution [22], without using any other prior knowledge (i.e., no semantic information). This scheme works well for natural images, especially scenes, because of large differences in their appearance. However, such differences are very subtle in histopathological images. For example, identifying subtle differences between benign and actionable categories may require characterizing cytoplasmic texture or nuclear appearance. This subtlety motivates us to leverage supervised information to design discriminative hash functions that are suitable for histopathological image retrieval.

3) Supervised Hashing: Intuitively, hashing methods minimize the Hamming distance of "neighboring" image pairs (e.g., close in terms of the Euclidean distance in the raw feature space). "Neighboring" in our case is defined by its semantic meaning, i.e., whether the two images belong to same category or not. Therefore, supervised information can be naturally encoded as similar and dissimilar pairs. Specifically, we assign the label 1 to image pairs when both are benign or actionable, and -1 to pairs when one is benign and the other is actionable, as shown in Fig. 3. l $(l \ll n)$ feature vectors are randomly selected from $\mathcal{X}$ to build the label matrix S. Note that we need to provide labels for only a small number of image pairs. Therefore, labeled data are explicitly constrained by both semantic information and visual similarities, whereas unlabeled data are mainly constrained by visual similarities and implicitly affected by labeled data.

Using this supervision scheme to bridge the semantic gap, r hash functions $h_k(\boldsymbol{x})_{k=1}^{r}$ are then designed to generate r discriminative hash bits based on Hamming distances. However, direct optimization of the following Hamming distances $\mathcal{D}_h(\boldsymbol{x}_i, \boldsymbol{x}_j) = |\{k | h_k(\boldsymbol{x}_i) \neq h_k(\boldsymbol{x}_j), 1 \leq k \leq r\}|$ is nontrivial. Therefore, code inner products can be used to simplify the

optimization process. As shown in [24], a Hamming distance and a code inner product are actually equivalent

$$\text{code}_r(\boldsymbol{x}_i) \circ \text{code}_r(\boldsymbol{x}_j) = r - 2\mathcal{D}_h(\boldsymbol{x}_i, \boldsymbol{x}_j) \tag{3}$$

where $\text{code}_r(\boldsymbol{x})$ are r-bit hash codes and the symbol $\circ$ is the code inner product.

Therefore, the objective function $\mathcal{Q}$ to the binary codes H_l is defined as

$$\min_{H_l \in \{1,-1\}^{l\times r}} \mathcal{Q} = \left\| \frac{1}{r} H_l H_l^\top - S \right\|_F^2 \tag{4}$$

where $H_l = \begin{bmatrix} h_1(\boldsymbol{x}_1), \ldots, h_r(\boldsymbol{x}_1) \\ \ldots\ldots \\ h_1(\boldsymbol{x}_l), \ldots, h_r(\boldsymbol{x}_l) \end{bmatrix}$ is the code matrix of the labeled data $\mathcal{X}_l$ and S is a label matrix with 1 for similar pairs and -1 for dissimilar pairs. $\|.\|_F$ denotes the Frobenius norm. Define $\bar{K}_l$ as $[\bar{\boldsymbol{k}}(\boldsymbol{x}_1), \ldots, \bar{\boldsymbol{k}}(\boldsymbol{x}_l)]^\top \in \mathbb{R}^{l\times m}$, $\bar{\boldsymbol{k}}(\boldsymbol{x}_i)$. The inner product of code matrix H_l can be represented as $H_l H_l^\top = \sum_{k=1}^{r} sgn(\bar{K}_l \boldsymbol{a}_k)(sgn(\bar{K}_l \boldsymbol{a}_k))^\top$ for binarization. Therefore, the new objective function $\mathcal{Q}$ that offers a clearer connection and easier access to the model parameter $\boldsymbol{a}_k$ is

$$\min_{\boldsymbol{a}_k} \quad \mathcal{Q}(\boldsymbol{a}_k) = \left\| \sum_{k=1}^{r} sgn(\bar{K}_l \boldsymbol{a}_k)(sgn(\bar{K}_l \boldsymbol{a}_k))^\top - rS \right\|_F^2. \tag{5}$$

C. Optimization and Online Updating

A greedy method is used for solving hash functions sequentially. A residue matrix is defined as $R_{k-1} = rS - \sum_{t=1}^{k-1} sgn(\bar{K}_l \boldsymbol{a}_t^\star)\ (sgn(\bar{K}_l \boldsymbol{a}_t^\star))^\top (R_0 = rS)$ when solving $\boldsymbol{a}_k^\star$, where $\boldsymbol{a}_t^\star$ is the previously solved vector. Using this residue matrix, the objective function becomes

$$\min_{\boldsymbol{a}_k} \quad g(\boldsymbol{a}_k) = -(sgn(\bar{K}_l \boldsymbol{a}_k))^\top R_{k-1} sgn(\bar{K}_l \boldsymbol{a}_k). \tag{6}$$

This objective function has two problems in generating hash functions accurately and efficiently. The first is that $g(\boldsymbol{a}_k)$ is neither convex nor smooth, making it hard to solve exactly. To alleviate this first problem, two approximation schemes can be employed to minimize g.

1) Spectral relaxation [21] is applied to drop the sign functions and hence convexify the object function

$$\max_{\boldsymbol{a}_k} \quad g(\boldsymbol{a}_k) = (\bar{K}_l \boldsymbol{a}_k)^\top R_{k-1} (\bar{K}_l \boldsymbol{a}_k) \quad s.t. (\bar{K}_l \boldsymbol{a}_k)^\top R_{k-1} (\bar{K}_l \boldsymbol{a}_k) = l. \tag{7}$$

   Although spectral relaxation results in fast optimization, it might deviate far from the optimal solution under a large l, as the relaxation removes the sign function.

2) Sigmoid smoothing is employed to replace $sgn()$ with the sigmoid-shaped function $\varphi(x) = 2/(1 + \exp(-x)) - 1$

$$\min_{\boldsymbol{a}_k} \quad \tilde{g}(\boldsymbol{a}_k) = -(\varphi(\bar{K}_l \boldsymbol{a}_k))^\top R_{k-1} \varphi(\bar{K}_l \boldsymbol{a}_k). \tag{8}$$

   Consequently, the objective function g can be minimized using the standard gradient-descent technique. Despite the accuracy, such a scheme is not efficient enough.

Algorithm 1: Optimization and Online Updating

Input: training feature vectors $\mathcal{X} = \{\boldsymbol{x}_i \in \mathbb{R}^d\}_{i=1}^n$, label matrix $S \in \mathbb{R}^{l\times l}$, old hash functions $h(\boldsymbol{x})$ if updating

initialize $R_0 = rS$ and $T_{max} = 500$;

for $k = 1; k \le r$ **do**

  if *is updating* **then**

    $\boldsymbol{a}_k^0 \leftarrow \boldsymbol{a}_k^\star$;

  end

  else

    solve the generalized eigenvalue problem $\bar{K}_l^\top R_{k-1} \bar{K}_l \boldsymbol{a} = \lambda \bar{K}_l^\top \bar{K}_l \boldsymbol{a}$ obtaining the largest eigenvector $\boldsymbol{a}_k^0$ such that $(\boldsymbol{a}_k^0)^\top \bar{K}_l^\top \bar{K}_l \boldsymbol{a}_k^0 = l$;

  end

  use the gradient descent method to optimize $min_a - (\varphi(\bar{K}_l \boldsymbol{a}_k))^\top R_{k-1} \varphi(\bar{K}_l \boldsymbol{a}_k)$ with the initial solution $\boldsymbol{a}_k^0$ and T_{max} budget iterations, achieving $\boldsymbol{a}_k^\star$;

  $h^0 \leftarrow sgn(\bar{K}_l \boldsymbol{a}_k^0) \quad h^\star \leftarrow sgn(\bar{K}_l \boldsymbol{a}_k^\star)$;

  if $(h^0)^\top R_{k-1} h^0 > (h^\star)^\top R_{k-1} h^\star$ **then**

    $\boldsymbol{a}_k^\star \leftarrow \boldsymbol{a}_k^0 \quad h^\star \leftarrow h^0$

  end

  $R_k = R_{k-1} - h^\star (h^s tar)^\top$;

end

Output: r hash functions $\{h_k(\boldsymbol{x}) = sgn(\bar{\boldsymbol{k}}^\top(\boldsymbol{x}) \boldsymbol{a}_k^\star)\}_{k=1}^r$

The second problem is how to efficiently update hash functions when new samples arrive. It is likely that not all training images are available at the beginning. When a new batch of labeled images come, regenerating hash functions on all training data is obviously not a good solution owing to the inefficiency. To solve these two above-mentioned issues, we propose a hybrid optimization scheme that combines both spectral relaxation and sigmoid smoothing, and also adapt the online updating concept to generate new hash functions efficiently with new samples. As shown in Algorithm 1, special relaxation is used for initialization purposes since it is very efficient. Then sigmoid smoothing is used for accurate generation of the hash functions. Given new data, instead of retraining the whole process we use previous results as the new initialization for sigmoid smoothing, i.e., as a warm start. Therefore, hashing functions can be updated on the fly. As shown in the experiments, this updating scheme is significantly faster than the traditional retraining process.

After the discriminative hash functions are obtained via optimizing g, high-dimensional image features can be mapped into informative binary bits. Specifically, when a query image comes, its extracted features are mapped into the kernel space and compressed as binary bits, using the hashing functions learned by our algorithm. Then, similar histopathological images in the training set can be retrieved by exhaustively but efficiently comparing their Hamming distances. Furthermore, these binary bits can be also indexed into a hash table for constant-time retrieval, independent of the number of images.

IV. Experiments

In this section, we discuss the experimental setting and results using microscopic breast-tissue images.

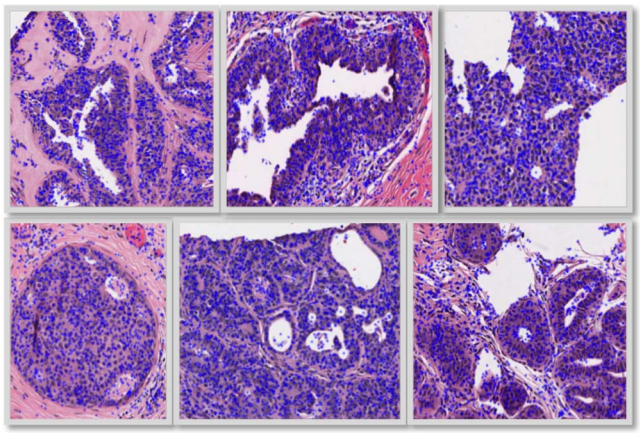

Fig. 4. Key points detected by difference of Gaussian (marked as blue stars).

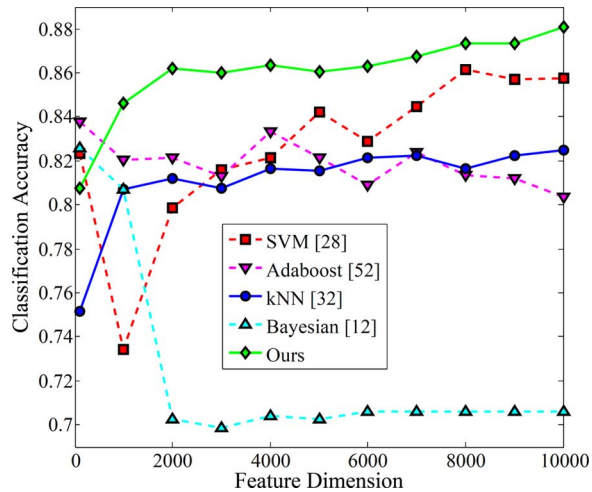

Fig. 5. Comparison of classification accuracy with different dimensions of features (from 100 to 10000).

A. Experimental Setting

Breast-tissue specimens available for this study were collected on a retrospective basis from the IU Health Pathology Lab (IUHPL) according to the protocol approved by the Institutional Review Board (IRB) for this study. All the slides were imaged using a ScanScope registered digitizer (Aperio, Vista, CA, USA) available in the tissue archival service at IUHPL. 3121 images (around 2250 K pixels) were sampled from 657 larger region-of-interest images (e.g., 5 K $\times$ 7 K) of microscopic breast tissue, which were gathered from 116 patients. 53 of these patients were labeled as benign (UDH) and 63 as actionable (ADH/DCIS), based on the majority diagnosis of nine board-certified pathologists. To demonstrate the efficiency of our method, one fourth of all patients in each category were randomly selected as the test set and the remainder used for training. Note that each patient may have different number of images. Therefore, the number of testing images is not fixed. The approximate number is about 700–900 in each testing process. All the experiments were conducted on a 3.40 GHz CPU with four cores and 16 G RAM, in a MATLAB implementation.

In each image, 1500–2000 SIFT descriptors were extracted from key points detected by DoG [50] (Fig. 4). These descriptors were quantized into sets of cluster centers using bag-of-words, in which the feature dimension equals the number of clusters. Specifically, we quantize them into high-dimensional feature vectors of length 10 000, to maximally utilize these millions of cell-level texture features. We provide both qualitative and quantitative evaluations for our proposed framework on two tasks, image classification (i.e., benign versus actionable category) and image retrieval, in terms of accuracy and computational efficiency.

In our system, classification is achieved using the majority vote of the top images retrieved by hashing. We compare our approach with various classifiers that have been widely used in systems for histopathological image analysis. Specifically, kNN has often been used as the baseline in analyzing histopathological images [32], [18], owing to its simplicity and proved lower bound, despite the inefficiency in large-scale databases. The Bayesian method is another solution to ensemble statistics of all extracted features and minimize the classification metric, which shows its efficacy in classifying histopathological images [12]. Boosting methods are always employed to combine multiple weak classifiers for higher accuracy [27], [18], [52]. SVM with a nonlinear kernel is commonly used in histopathological images because of its efficiency and the ability to handle linearly inseparable cases [53], [28], [54], [9]. For fair comparison, all parameters and kernel selections of these compared methods were optimized by cross-validation.

In addition, we also compared our proposed method with several dimensionality-reduction algorithms in terms of classification accuracy. Principal component analysis (PCA) has been widely used in this area to preserve variance of original features [55]. Graph embedding is a nonlinear dimensionality-reduction algorithm that performs well in grading of lymphocytic infiltration in HER2+ breast cancer histopathology [30]. Since we use supervised information in generating hash functions, a supervised dimensionality reduction algorithm, neighborhood components analysis (NCA) [56], was also chosen for our experimental comparisons.

B. Evaluation of Image Classification

To demonstrate the benefit of high-dimensional features, all methods are evaluated on multiple dimensions of SIFT quantization, ranging from 100 to 10 ,000. In our proposed framework, we use a hashing method to compress all features to 48 bits (only six bytes), which is a suitable size for using a hash table. To compare with dimensionality-reduction methods, we compress all features into 48 dimensions. In contrast to hashing methods using bits, results of other methods are based on floats or doubles. Therefore, they need much more storage room than hashing results do.

Fig. 5 shows the quantitative results for the classification accuracy. Most methods achieve better accuracy with higher-dimensional features. This is very intuitive, as finer quantization of SIFT features usually provides richer information. In particular, since the SIFT interest points cover most nuclear regions

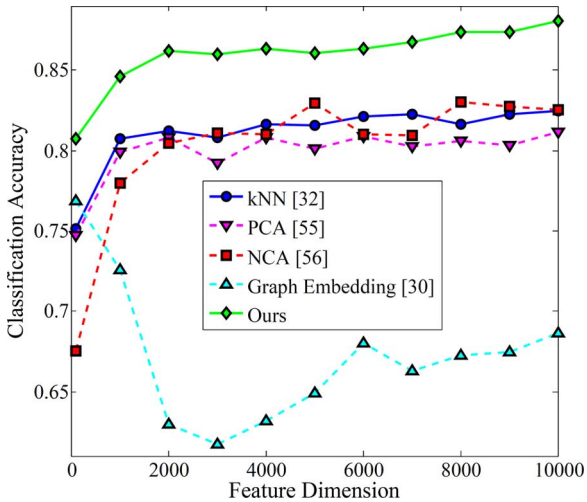

Fig. 6. Comparison of the classification accuracy for different dimensionality-reduction methods.

in images, fine quantization (i.e., high-dimensional features) indicates analysis on a small scale. Exceptions are the Adaboost and Bayestian methods, whose accuracy drops when the feature dimensions increase. This indicates that high-dimensional features do not guarantee the improvement of accuracy. An important factor is the proper utilization of such information. For example, Adaboost is essentially a feature-selection method that chooses only an effective subset of features for the classification. Therefore, it may lose important information, especially in high-dimensional space, resulting in accuracy worse than that of our hashing method. Our method is also generally better than kNN and its variations, owing to the semantic information (i.e., labels of similar and dissimilar pairs in hashing) that bridges the semantic gap between images and diagnoses. Note that our hashing method needs only a small amount of supervision—in this case, similar or dissimilar pairs of 40% images. This is generally less than the supervised information required by SVM in the training stage. It compares favorably to all other methods when the feature dimension is larger than 1000. The overall classification accuracy is 88.1% for 10 000-dimensional features, 2%–18% better than other methods.

Fig. 6 shows the comparison with various dimensionality-reduction methods that have been used in analyzing histopathological images, including PCA [55], graph embedding [30], and NCA [56]. Our approach consistently outperforms the other methods. In particular, graph embedding, a nonlinear dimensionality-reduction method, fails to handle this challenging data set, especially in high-dimensional space, probably because of over-fitting and inappropriate assumption of the nonlinearity. Therefore, its results are worse than PCA for this data set. NCA is slightly better than PCA, owing to the supervised information during compression. Compared to kNN results that use uncompressed features, PCA is slightly worse and NCA is comparable in general. However, both of them are 5% worse than our method.

Fig. 7 compares the computational efficiency of these methods. With increasing dimensionality the running time of some compared methods increases dramatically. When feature dimensionality reaches 10 000, kNN needs 16 s to classify all query images, and Adaboost needs 5 s. SVM, dimensionality-reduction methods, and the proposed method are much faster. However, the running time for SVM increases with the feature dimensionality, as shown in the expanded view of Fig. 7. In contrast, PCA, graph embedding, NCA, and ours achieve constant running time in this data set owing to the fixed size of features after compression. Compared to other dimensionality-reduction methods, our approach is about 10 times faster because of the efficient comparison among binary codes. In addition, the running time of all kNN-based methods increases with the number of images in a data set, as exhaustive search is needed, while hashing-based methods can achieve $\mathcal{O}(1)$ efficiency using a hash table. To summarize, the average running time of our method is merely 0.01 s for all testing images, which is 40 times faster than SVM and 1500 times faster than kNN.

The online updating scheme was also evaluated and compared with the traditional retraining scheme. Specifically, the running time was evaluated for both the online updating and offline retraining schemes, using different numbers of new training samples ranging from 1% to 20%. As shown in Fig. 8, the online updating scheme is three to four times faster than the offline retraining, indicating better scalability and efficiency. This is expected, since our online updating scheme focuses mainly on analyzing the new data, which are a small portion of the whole data set. In fact, when the existing database is larger, the improvement will be even more significant. We have also conducted experiments to compare the classification accuracy of the online updating approach and the offline-retraining scheme. When 1%–20% new samples are added, the classification accuracy of both methods is generally comparable. On average, the retraining process is 0.035% more accurate than our online updating, which is negligible compared to around 88% overall accuracy. This demonstrates that our method can update the hashing model efficiently without sacrificing classification accuracy In fact, handling 20% new samples is already adequate for practical use, since new data usually come gradually and each model updating process needs to incorporate just a relatively small portion.

C. Evaluation of Image Retrieval

We have also conducted experiments on image retrieval using 10 000-dimensional features. The retrieval precision, evaluated at a given cut-off rank and considering only the topmost results, is reported in Table I, along with the query time and memory cost. The results are quite consistent with the image classification. The mean precision of the hashing method is around 83%, and the standard deviation is 1.1%, which is much better than PCA [55], graph embedding [30] and NCA [56]. In most cases, the precision of our method is at least 6% better than the others, except the NCA. Our method is around 3.5% better than NCA on benign cases. To demonstrate statistical significance, we perform $t\text{-}test$ for the precision obtained by NCA and by the proposed method on benign cases, under the null hypothesis using a significance level of 0.05. The $p\text{-}values$ are found as

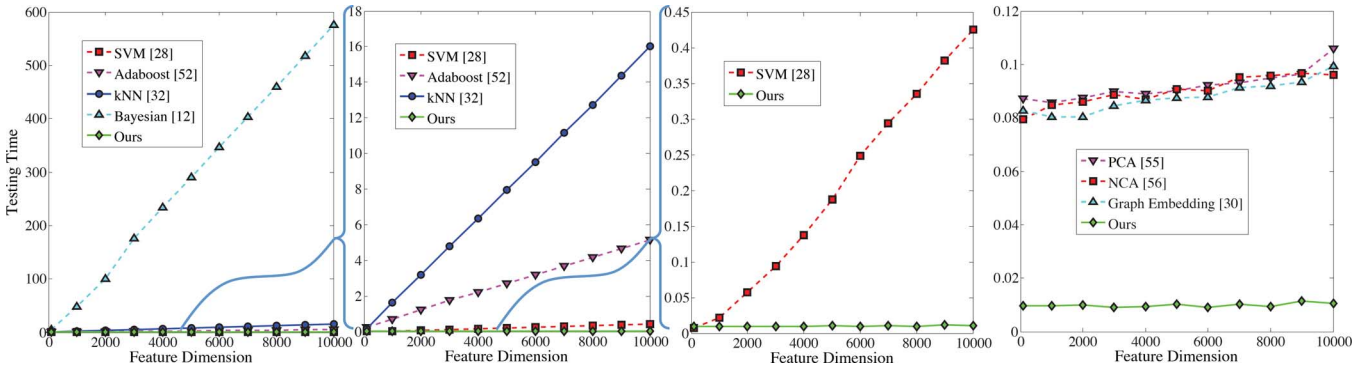

Fig. 7. Comparison of the classification running time (seconds) with different dimensions of features, which means the average time of classifying hundreds of test images.

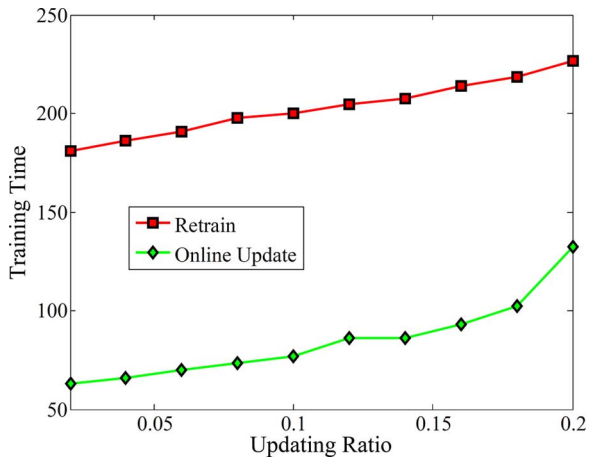

Fig. 8. Comparison of the training time (seconds) using different optimization schemes.

3.6×10^{-6}, 3.2×10^{-6}, and 5.7×10^{-6} at the range of top 10, 20, and 30 retrievals, respectively, demonstrating that precision values achieved by the proposed technique are indeed significantly better than NCA on the benign cases. In addition, our method is around 14% better than NCA in the actionable cases, resulting much higher average precision. In fact, most traditional methods produce such highly unbalanced results as NCA does, i.e., the retrieval precision of the benign category is much higher than that of the actionable one. In contrast, our method does not have this problem, owing to the supervised information and the optimization for balanced hash bits. Our framework is also computationally more efficient than traditional methods. The query time of our hashing method is a thousand times faster than kNN and ten times faster than other dimensionality-reduction methods. Note that our method takes a constant time when using the hash table, independent of the number of feature dimensions and the number of samples. Furthermore, the memory cost is also considerably reduced (10 000 times less than that of kNN). Therefore, this method is more applicable to large-scale databases (millions of images) than are other methods.

Fig. 9 shows our image-retrieval results. The top five relevant images are listed for each query image. The differences between certain images in different categories are very subtle. Our accurate results demonstrate the efficacy of the proposed method. Specifically, the features capturing local texture and appearance are very robust to various image sizes, cell distributions, and occlusions by the blood. The supervised information also improves the retrieval precision by correlating binary code with diagnosis information. These retrieved images are clinically relevant in potential (i.e., retrieved images belong to the same category as the query image) and thus can be useful for decision support.

D. Discussions

We discuss the benefits of the algorithm, parameter sensitivity, implementation issues, and limitations here.

Regarding the choice of high-dimensional features, around 1000 dimensions have usually been used for quantization by many previous studies, a number that has been proved to achieve good accuracy. Using lower-dimensional features (e.g., 100) is not accurate, while using higher-dimensional features is not efficient, and the improvement of accuracy could be marginal. This is consistent with our experimental results shown in Fig. 5, i.e., a performance jump from 100 to 1000 dimensions. On the other hand, when analyzing histopathological images, using high-dimensional features (e.g., 10 000) implies nearly cell-level analysis, which is actually beneficial for the accuracy, even though the accuracy gain is not as big as jumping from 100 to 1000. Therefore, we have introduced hashing methods to harvest the benefits of high-dimensional features, without sacrificing computational efficiency.

Regarding supervised information, it significantly improves classification accuracy thanks to the discriminative modeling of the hashing function in an attempt to bridge the semantic gap. In Fig. 10, we randomly selected 100 samples from benign and actionable categories and visualized their 48 hash bits. The distributions of hash bits are clearly different between the two categories, explaining the high accuracy for classification. We also quantitatively investigated the benefits of using supervised information. Specifically, we evaluated our method when using 10% to 100% supervision or training labels, as shown in Fig. 11. The gain in accuracy is very high (from 71% to nearly 87%) when the ratio of training labels increases from 10% to 40%, which demonstrates the efficacy of using supervised information. For more than 40% labels, the improvement of accuracy becomes marginal, reaching 88% accuracy when using 100%

TABLE I
COMPARISON OF RETRIEVAL PRECISION FOR THE TOP 10, 20, AND 30 RESULTS, ALONG WITH THE MEMORY COST OF TRAINING DATA AND QUERY TIME OF ALL TEST IMAGES. BOTH MEAN VALUES AND THE STANDARD DEVIATION (STD) OF 20 EXPERIMENTS ARE REPORTED. BEST PRECISION IN EACH ROW FOR BENIGN AND ACTIONABLE CATEGORIES ARE HIGHLIGHTED

	kNN [32]		PCA [55]		NCA [56]		Graph Embedding [30]		Ours	
	benign	actionable	benign	actionable	benign	actionable	benign	actionable	benign	actionable
P@10	0.779	0.687	0.762	0.705	0.799	0.697	0.672	0.487	**0.836**	**0.830**
P@20	0.773	0.653	0.758	0.681	0.800	0.689	0.673	0.486	**0.839**	**0.829**
P@30	0.770	0.631	0.755	0.667	0.800	0.685	0.670	0.480	**0.837**	**0.833**
STD	0.024		0.028		0.020		0.012		0.011	
Time (s)	15.77		10.07		10.04		10.03		<0.01	
Memory	134.58MB		0.65MB		0.65MB		0.65MB		0.01MB	

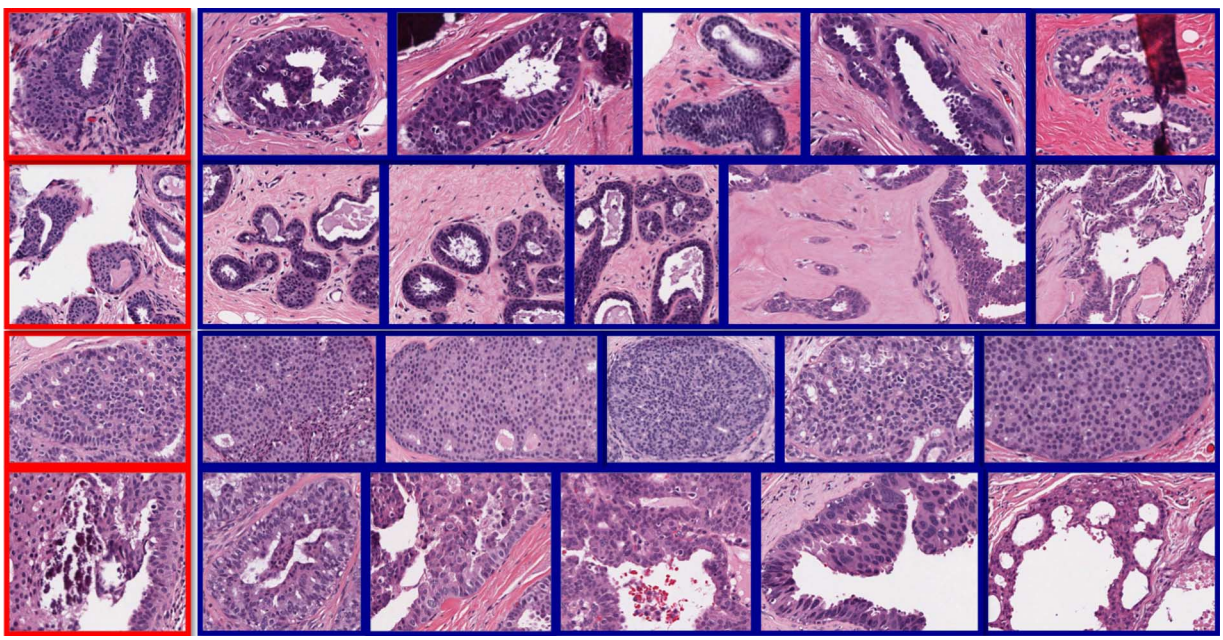

Fig. 9. Four examples of our image retrieval (query marked in red, and retrieved images marked in blue). First two rows are benign; the last two rows are actionable.

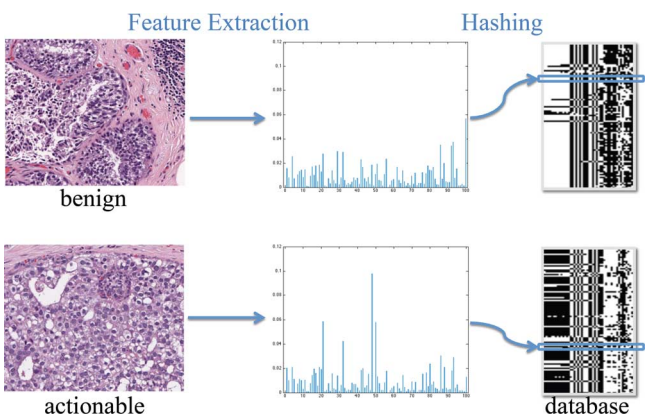

Fig. 10. Visualization of compressed hash bits. Their distribution well separates the begin and actionable categories.

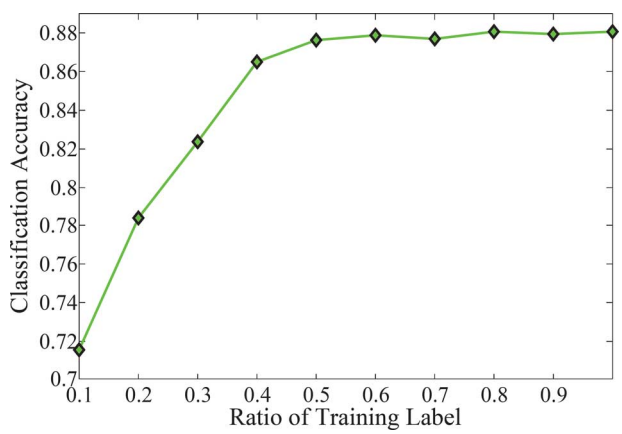

Fig. 11. Classification accuracy when using 10%–100% supervision.

labels. This means that our method needs only a small portion of labels to achieve high accuracy, owing to the unified framework of coupling Hamming distance optimization and supervised information.

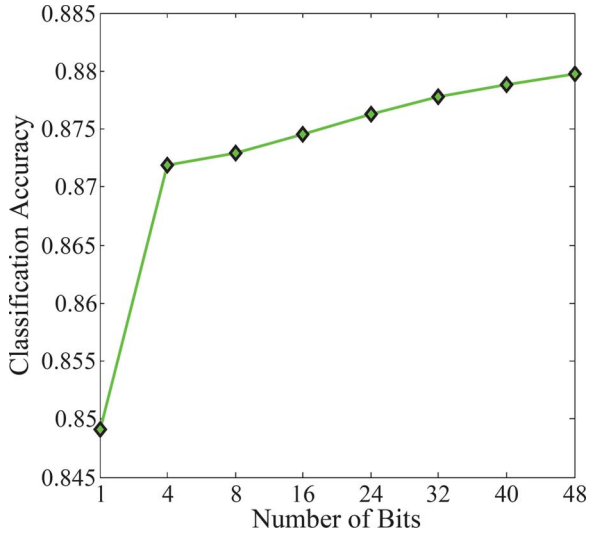

Fig. 12. Classification accuracy with different lengths of hashing bits.

One of the most significant benefits of our proposed framework is the computational and storage efficiency. Comparing 48 bits with Hamming distance or hash table is substantially faster than using high-dimensional features. However, a natural question is whether the length of hashing bits affects the accuracy and retrieval precision. Therefore, we evaluated the effect of hashing-bit lengths ranging from 1 to 48. Theoretically, 1 bit is sufficient for binary classification purpose, i.e., actionable versus benign. In fact, as shown in Fig. 12, using 8 bits already achieves high accuracy for classification. However, such short code is not discriminative enough for image retrieval. For example, 8 bits can represent only 64 hash values. This means that nearly 50 images are mapped into the same hash value, which is an unordered list with zero Hamming distance. Retrieving them may not be beneficial for decision support. On the other hand, using more than 64 bits adversely affects computational efficiency, since the hash table is no longer an option owing to memory constraint. Therefore, we chose 48 bits for this task, ensuring sound accuracy for classification and high relevance for retrieval without sacrificing efficiency. We expect that our scalable framework can be efficiently used for real-time querying of very large databases.

In the task of image retrieval, our method effectively retrieves images with morphological and architectural image patterns similar to the query image, as shown in Fig. 9. This can be explained by the capability of the hashing function in leveraging both diagnostic information and visual similarities. In other words, hash bits can simultaneously encode local textural features with semantic labels.

V. Conclusion

In this paper, we developed a *scalable image-retrieval framework* for intelligent histopathological image analysis. Specifically, we employed hashing to achieve efficient image retrieval and presented an improved kernelized and supervised hashing approach for real-time image retrieval. The potential applications of our framework include image-guided diagnosis, decision support, education, and efficient data management. In our future work, we will examine more types of features, especially those features stemming from segmentation and architectures. Furthermore, we will incorporate appropriate feature-fusion techniques to design a hybrid hashing method such that multiple types of features can be systematically fused to boost image-retrieval accuracy. We will also evaluate our framework in more applications in histopathological image analysis.

References

[1] R. Siegel, D. Naishadham, and A. Jemal, "Cancer statistics, 2013," *CA, Cancer J. Clinicians*, vol. 63, no. 1, pp. 11–30, 2013.
[2] D. C. Sgroi, "Preinvasive breast cancer," *Annu. Rev. Pathol. Mech. Disease*, vol. 5, pp. 193–221, 2010.
[3] D. L. Page, W. D. Dupont, L. W. Rogers, and M. S. Rados, "Atypical hyperplastic lesions of the female breast. A long-term follow-up study," *Cancer*, vol. 55, no. 11, pp. 2698–2708, 1985.
[4] B. Weyn, G. v. d. Wouwer, A. v. Daele, P. Scheunders, D. v. Dyck, E. v. Marck, and W. Jacob, "Automated breast tumor diagnosis and grading based on wavelet chromatin texture description," *Cytometry*, vol. 33, no. 1, pp. 32–40, 1998.
[5] C. Bilgin, C. Demir, C. Nagi, and B. Yener, "Cell-graph mining for breast tissue modeling and classification," in *Proc. IEEE Int. Conf. Eng. Med. Biol. Soc.*, 2007, pp. 5311–5314.
[6] M. N. Gurcan, L. E. Boucheron, A. Can, A. Madabhushi, N. M. Rajpoot, and B. Yener, "Histopathological image analysis, A review," *IEEE Rev. Biomed. Eng.*, vol. 2, pp. 147–171, Dec. 2009.
[7] J. Kong, O. Sertel, H. Shimada, K. L. Boyer, J. H. Saltz, and M. N. Gurcan, "Computer-aided evaluation of neuroblastoma on whole-slide histology images: Classifying grade of neuroblastic differentiation," *Pattern Recognit.*, vol. 42, no. 6, pp. 1080–1092, 2009.
[8] P.-W. Huang and C.-H. Lee, "Automatic classification for pathological prostate images based on fractal analysis," *IEEE Trans. Med. Imag.*, vol. 28, no. 7, pp. 1037–1050, Jul. 2009.
[9] P.-W. Huang and Y.-H. Lai, "Effective segmentation and classification for HCC biopsy images," *Pattern Recognit.*, vol. 43, no. 4, pp. 1550–1563, 2010.
[10] P. Khurd, C. Bahlmann, P. Maday, A. Kamen, S. Gibbs-Strauss, E. M. Genega, and J. V. Frangioni, "Computer-aided Gleason grading of prostate cancer histopathological images using texton forests," in *Proc. IEEE Int. Symp. Biomed. Imag.*, 2010, pp. 636–639.
[11] M. Veta, J. Pluim, P. v. Diest, and M. Viergever, "Breast cancer histopathology image analysis: A review," *IEEE Trans. Biomed. Eng.*, vol. 61, no. 5, pp. 1400–1411, May 2014.
[12] D. Comaniciu, P. Meer, and D. J. Foran, "Image-guided decision support system for pathology," *Mach. Vis. Appl.*, vol. 11, no. 4, pp. 213–224, 1999.
[13] L. Zheng, A. W. Wetzel, J. Gilbertson, and M. J. Becich, "Design and analysis of a content-based pathology image retrieval system," *IEEE Trans. Inf. Technol. Biomed.*, vol. 7, no. 4, pp. 249–255, Dec. 2003.
[14] H. Müller, N. Michoux, D. Bandon, and A. Geissbuhler, "A review of content-based image retrieval systems in medical applications-clinical benefits and future directions," *Int. J. Med. Inf.*, vol. 73, no. 1, pp. 1–23, 2004.
[15] H. C. Akakin and M. N. Gurcan, "Content-based microscopic image retrieval system for multi-image queries," *IEEE Trans. Inf. Technol. Biomed.*, vol. 16, no. 4, pp. 758–769, Apr. 2012.
[16] 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. Digital Imag.*, vol. 26, no. 6, pp. 1025–1039, 2013.
[17] T. M. Lehmann, M. O. Güld, T. Deselaers, D. Keysers, H. Schubert, K. Spitzer, H. Ney, and B. B. Wein, "Automatic categorization of medical images for content-based retrieval and data mining," *Comput. Med. Imag. Graph.*, vol. 29, no. 2, pp. 143–155, 2005.
[18] L. Yang, W. Chen, P. Meer, G. Salaru, L. A. Goodell, V. Berstis, and D. J. Foran, "Virtual microscopy and grid-enabled decision support for large-scale analysis of imaged pathology specimens," *IEEE Trans. Inf. Technol. Biomed.*, vol. 13, no. 4, pp. 636–644, Jul. 2009.
[19] D. J. Foran *et al.*, "Imageminer, A software system for comparative analysis of tissue microarrays using content-based image retrieval, high-performance computing, and grid technology," *J. Am. Med. Informat. Assoc.*, vol. 18, no. 4, pp. 403–415, 2011.
[20] M. Datar, N. Immorlica, P. Indyk, and V. S. Mirrokni, "Locality-sensitive hashing scheme based on p-stable distributions," in *Proc. Annu. ACM Symp. Computat. Geometry*, 2004, pp. 253–262.

[21] Y. Weiss, A. Torralba, and R. Fergus, "Spectral hashing," in *Adv. Neural Inf. Process. Syst.*, 2008, pp. 1753–1760.

[22] B. Kulis and K. Grauman, "Kernelized locality-sensitive hashing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 6, pp. 1092–1104, Jun. 2012.

[23] J. Wang, S. Kumar, and S.-F. Chang, "Semi-supervised hashing for scalable image retrieval," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 3424–3431.

[24] W. Liu, J. Wang, R. Ji, Y.-G. Jiang, and S.-F. Chang, "Supervised hashing with kernels," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 2074–2081.

[25] X. Zhang, W. Liu, and S. Zhang, "Mining histopathological images via hashing-based scalable image retrieval," in *Proc. IEEE Int. Symp. Biomed. Imag.*, 2014, pp. 1111–1114.

[26] S. Petushi, F. U. Garcia, M. M. Haber, C. Katsinis, and A. Tozeren, "Large-scale computations on histology images reveal grade-differentiating parameters for breast cancer," *BMC Med. Imag.*, vol. 6, no. 1, p. 14, 2006.

[27] L. Yang, W. Chen, P. Meer, G. Salaru, M. D. Feldman, and D. J. Foran, "High throughput analysis of breast cancer specimens on the grid," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervent.*, 2007, pp. 617–625.

[28] J. C. Caicedo, A. Cruz, and F. A. Gonzalez, "Histopathology image classification using bag of features and kernel functions," *Artif. Intell. Med.*, pp. 126–135, 2009.

[29] 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.

[30] A. N. Basavanhally, S. Ganesan, S. Agner, J. P. Monaco, M. D. Feldman, J. E. Tomaszewski, G. Bhanot, and A. Madabhushi, "Computerized image-based detection and grading of lymphocytic infiltration in HER2+ breast cancer histopathology," *IEEE Trans. Biomed. Eng.*, vol. 57, no. 3, pp. 642–653, Mar. 2010.

[31] M. M. Dundar, S. Badve, G. Bilgin, V. Raykar, R. Jain, O. Sertel, and M. N. Gurcan, "Computerized classification of intraductal breast lesions using histopathological images," *IEEE Trans. Biomed. Eng.*, vol. 58, no. 7, pp. 1977–1984, Jul. 2011.

[32] A. Tabesh, M. Teverovskiy, H.-Y. Pang, V. P. Kumar, D. Verbel, A. Kotsianti, and O. Saidi, "Multifeature prostate cancer diagnosis and Gleason grading of histological images," *IEEE Trans. Med. Imag.*, vol. 26, no. 10, pp. 1366–1378, Oct. 2007.

[33] S. Doyle, S. Agner, A. Madabhushi, M. Feldman, and J. Tomaszewski, "Automated grading of breast cancer histopathology using spectral clustering with textural and architectural image features," in *Proc. IEEE Int. Symp. Biomed. Imag.*, 2008, pp. 496–499.

[34] J. G. Dy, C. E. Brodley, A. Kak, L. S. Broderick, and A. M. Aisen, "Unsupervised feature selection applied to content-based retrieval of lung images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 3, pp. 373–378, Feb. 2003.

[35] 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.

[36] H. Greenspan and A. T. Pinhas, "Medical image categorization and retrieval for PACS using the GMM-KL framework," *IEEE Trans. Inf. Technol. Biomed.*, vol. 11, no. 2, pp. 190–202, Mar. 2007.

[37] Y. Song, W. Cai, and D. Feng, "Hierarchical spatial matching for medical image retrieval," in *ACM Int. Workshop Med. Multimedia Anal. Retrieval*, 2011, pp. 1–6.

[38] F. Schnorrenberg, C. Pattichis, C. Schizas, and K. Kyriacou, "Content-based retrieval of breast cancer biopsy slides," *Technol. Health Care*, vol. 8, no. 5, pp. 291–297, 2000.

[39] X. S. Zhou, S. Zillner, M. Moeller, M. Sintek, Y. Zhan, A. Krishnan, and A. Gupta, "Semantics and CBIR: A medical imaging perspective," in *ACM Int. Conf. Content-Based Image Video Retrieval*, 2008, pp. 571–580.

[40] H. Müller, A. Geissbühler, and P. Ruch, "ImageCLEF 2004: Combining image and multi-lingual search for medical image retrieval," in *Multilingual Inform. Access for Text, Speech and Images*. New York: Springer, 2005, pp. 718–727.

[41] H. Müller and J. Kalpathy-Cramer, "The imageCLEF medical retrieval task at ICPR 2010–Information fusion," in *Proc. IEEE Int. Conf. Pattern Recognit.*, 2010, pp. 3284–3287.

[42] L. Yang, R. Jin, L. Mummert, R. Sukthankar, A. Goode, B. Zheng, S. C. Hoi, and M. Satyanarayanan, "A boosting framework for visuality-preserving distance metric learning and its application to medical image retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 1, pp. 30–44, Jan. 2010.

[43] G. Langs, H. Müller, B. H. Menze, and A. Hanbury, "VISCERAL: Towards large data in medical imaging—Challenges and directions," in *MCBR-CDS MICCAI Workshop*, 2013, vol. 7723.

[44] A. Hanbury, H. Müller, G. Langs, and B. H. Menze, "Cloud-based evaluation framework for big data," in *FIA Book 2013,*. New York: Springer, 2013, LNCS.

[45] Y. Mu, J. Shen, and S. Yan, "Weakly-supervised hashing in kernel space," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 3344–3351.

[46] J. Wang, S. Kumar, and S.-F. Chang, "Semi-supervised hashing for large-scale search," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 12, pp. 2393–2406, Dec. 2012.

[47] X. Liu, J. He, D. Liu, and B. Lang, "Compact kernel hashing with multiple features," in *Proc. ACM Int. Conf. Multimedia*, 2012, pp. 881–884.

[48] B. Kulis and T. Darrell, "Learning to hash with binary reconstructive embeddings," in *Adv. Neural Inf. Process. Syst.*, 2009, pp. 1042–1050.

[49] M. Norouzi and D. M. Blei, "Minimal loss hashing for compact binary codes," in *Int. Conf. on Mach. Learning*, 2011, pp. 353–360.

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

[51] J. Wang *et al.*, "Bag-of-features based medical image retrieval via multiple assignment and visual words weighting," *IEEE Trans. Med. Imag.*, vol. 30, no. 11, pp. 1996–2011, Nov. 2011.

[52] S. Doyle, M. Feldman, J. Tomaszewski, and A. Madabhushi, "A boosted Bayesian multiresolution classifier for prostate cancer detection from digitized needle biopsies," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 5, pp. 1205–1218, May 2012.

[53] O. Tuzel, L. Yang, P. Meer, and D. J. Foran, "Classification of hematologic malignancies using texton signatures," *Pattern Anal. Appl.*, vol. 10, no. 4, pp. 277–290, 2007.

[54] K. Nguyen, A. K. Jain, and R. L. Allen, "Automated gland segmentation and classification for Gleason grading of prostate tissue images," in *Proc. IEEE Int. Conf. Pattern Recognit.*, 2010, pp. 1497–1500.

[55] O. Sertel, J. Kong, U. V. Catalyurek, G. Lozanski, J. H. Saltz, and M. N. Gurcan, "Histopathological image analysis using model-based intermediate representations and color texture: Follicular lymphoma grading," *J. Signal Process. Syst.*, vol. 55, no. 1–3, pp. 169–183, 2009.

[56] J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov, "Neighbourhood components analysis," *Adv. Neural Inf. Process. Syst.*, 2004.