MINING HISTOPATHOLOGICAL IMAGES VIA HASHING-BASED SCALABLE IMAGE RETRIEVAL

Xiaofan Zhang[1], *Wei Liu*[2], *Shaoting Zhang*[1]

[1]Department of Computer Science, University of North Carolina at Charlotte, NC, USA
[2]IBM T. J. Watson Research Center, NY, USA

ABSTRACT

Automatic analysis of histopathological images has been widely investigated using computational image processing and machine learning techniques. Computer-aided diagnosis (CAD) systems and content-based image retrieval (CBIR) systems have been successfully developed for diagnosis, disease detection, and decision support in this area. In this paper, we focus on a *scalable* image retrieval method with high-dimensional features for the analysis of histopathology images. Specifically, we present a kernelized and supervised hashing method. With a small amount of supervised information, our method can compress a 10,000-dimensional image feature vector into only tens of binary bits with informative signatures preserved, and these binary codes are then indexed into a hash table that enables real-time retrieval. We validate the hashing-based image retrieval framework on several thousands of images of breast microscopic tissues for both image classification (i.e., benign vs. actionable categorization) and retrieval. Our framework achieves high search accuracy and promising computational efficiency, comparing favorably with other commonly used methods.

Index Terms— histopathological image analysis, breast lesion, CBIR, scalable image retrieval, hashing

1. INTRODUCTION

Breast cancer is the second most common cancer in the United States [13]. Fortunately, early detection with percutaneous biopsy can significantly increase the survival rates of patients. The usual ductal hyperplasia (UDH), atypical ductal hyperplasia (ADH) and ductal carcinoma in situ (DCIS) are the three stages in the development procedure from a normal terminal duct-lobular unit to an invasive cancer. Each stage has a higher risk to develop into invasive breast carcinoma [12]. Therefore, the therapy planning and management relies on the diagnosis of UDH and ADH/DCIS. However, classifying these stages is inexact and depends on subjective assessment of the pathologists, which poses a special challenge in the diagnosis of pre-invasive breast cancer.

Computer-aided diagnosis (CAD) systems have been employed for reliable and consistent identification of these stages, using high-resolution images digitized from tissue histopathology slides [6]. For examples, Petushi, et al. [11] proposed to identify cell nuclei in histopathology slide images and classify them in a supervised classification scheme according to morphology. Doyle et al. [3] used support vector machine (SVM) with texture-based and nuclear architecture-based features to distinguish between cancerous and non-cancerous cases, and predict the grades of the breast cancer. Dundar et al. [4] proposed a binary classifier using size, shape, and intensity-based features extracted from identified cells, which achieved promising accuracy.

Besides classifier-based CAD systems, content-based image retrieval (CBIR) has also been widely investigated for decision support in digital pathology and many other clinical applications [5, 10]. Given an image database with ground truth recorded, CBIR methods aim to retrieve and display images with morphological profiles most relevant and consistent to the query image. The retrieved images also indicate the most likely diagnosis (e.g., classification results) using majority logic.

Despite the efficacy of existing CBIR systems, new opportunities and challenges arise with the ever-increasing 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, it is challenging to maintain the retrieval efficiency with such large-scale data and high-dimensional features. Although cloud- and grid-computing is a potential solution [5, 16], few efforts have been put on the computational and scalable algorithms in this area.

In this paper, we focus on the scalable image retrieval methods for the image-guided diagnosis of pre-invasive breast cancer. Particularly, we investigate hashing-based methods [2, 15, 7, 8] for scalable and high-dimensional image retrieval. A kernel-based supervised hashing model is introduced. With a small amount of labeled information, it is able to encode a high-dimensional image feature vector to short binary codes. Such compact code has enabled significant efficiency gains in the storage. It also allows real-time search even in a collection of millions of images, owing to the hash table of binary code. We validate this proposed method on several thousands of breast tissue images. The experimental results demonstrate the accuracy and efficiency of our framework.

978-1-4673-1961-4/14/$31.00

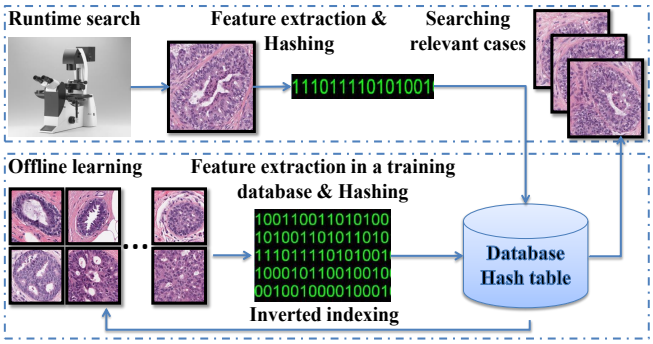

Fig. 1. Overview of our proposed system.

2. METHODOLOGY

2.1. Overview of Scalable Image Retrieval Framework

Fig.1 shows our proposed framework of scalable image retrieval-based diagnosis system. In offline learning, we first extract high-dimensional features of the texture and appearance from digitized histopathological images based on SIFT [9] and bag-of-words [14]. These effective features have been used in both general computer vision tasks and histopathological image analysis [1]. Although these features can be directly used to measure the difference between image pairs, computational efficiency is an issue, especially when searching in a large database (i.e., searching k-nearest neighbors exhaustively). Therefore, we employ hashing method to compress these features into binary codes with tens of bits. Such short binary features allow mapping easily into a hash table for real-time search. Each feature is then linked to the corresponding training images using inverted index. During runtime query, high-dimensional features are extracted from the query image and then projected to the binary codes. With hash table, searching nearest neighbors is in constant time, no matter 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 using majority voting.

2.2. Kernelized and Supervised Hashing

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

Hashing Method: Given a set of image feature vectors $\mathcal{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_n\} \subset \mathbb{R}^d$ (in our case, $\mathbf{x}_i$ is the high-dimensional SIFT feature vector 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. 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 efficient Hamming distance computations. Note that hashing method is different from dimension reduction, since it needs to ensure that the generated hash bits have balanced and uncorrelated bit distributions, which leads to maximum information at each single bit and minimum redundancies among all bits.

Kernelized Hashing: Kernel methods can handle practical data that is mostly linearly inseparable. For histopathological images, the phenomena of linear inseparability really happen. Therefore, kernel functions should be considered in hashing methods $h = sign(f(x))$ [7]. A kernel function is denoted as κ: $\mathbb{R}^d \times \mathbb{R}^d \mapsto \mathbb{R}$. The prediction function f: $\mathbb{R}^d \mapsto \mathbb{R}$ with kernel κ plugged in is defined as:

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

where $\mathbf{x}_{(1)}, \cdots, \mathbf{x}_{(m)}$ are $m (m \ll n)$ random samples 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 should keep as much information as possible, i.e., $\sum_{i=1}^{n} h(\mathbf{x}_i) = 0$. Therefore, b is set as the median of $\{\sum_{j=1}^{m} \kappa(\mathbf{x}_{(j)}, \mathbf{x}_i) a_j\}_{i=1}^{n}$, which is usually approximated by the mean. Adding this constraint into Eq. 1, we obtain

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

Denote $\mathbf{a} = [a_1, a_2, ..., a_m]^T$. $\mathbf{a}$ is the most important factor that determines hash functions, In traditional kernelized hashing methods, $\mathbf{a}$ is defined as a random direction drawn from a Gaussian distribution [7], without using any supervised information. This scheme works well for natural images, especially scene images, because of large differences in their appearances. However, such differences are very subtle in histopathological images. This motivates us to leverage supervised information to design discriminative hash functions that are suitable for histopathological image retrieval.

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). Therefore, supervised information can be naturally encoded as similar and dissimilar pairs. Specifically, we assign label 1 to image pairs when both are benign or actionable, and -1 to pairs when one is benign and the other is actionable. Note that we only need to provide labels for a small amount of image pairs. The undefined image pairs are labeled as 0. Using such supervision, r hash functions $h_k(\mathbf{x})_{k=1}^{r}$ are then designed to generate r discriminative hash bits based on Hamming distances. However, directly optimizing the following Hamming distances is complex: $\mathcal{D}_h(\mathbf{x}_i, \mathbf{x}_j) = |\{k | h_k(\mathbf{x}_i) \neq h_k(\mathbf{x}_j), 1 \leq k \leq r\}|$. Therefore, code inner products can be used to simplify the optimization process. As shown in [8], a Hamming distance and a code inner product are actually equivalent. The least-squares style objective function $\mathcal{Q}$ to the binary codes H_l is:

$$\min_{H_l \in \{1,-1\}^{l \times r}} \mathcal{Q} = \left\| \frac{1}{r} H_l H_l^T - S \right\|_F^2, \tag{3}$$

where H_l is the the code matrix of the labeled data $\mathcal{X}_l$, S is a label matrix consisting of 1 for similar pairs, -1 for dissimilar pairs, and 0 for undefined pairs. $\|.\|_F$ denotes the Frobenius norm. The code matrix H_l is represented as $H_l = sgn(\bar{K}_l A)$ for binarization, where $\bar{K}_l = [\bar{\mathbf{k}}(\mathbf{x}_1), \cdots, \bar{\mathbf{k}}(\mathbf{x}_l)]^T \in \mathbb{R}^{l \times m}$, $\bar{\mathbf{k}}(\mathbf{x}_i)$ is a kernelized vectorial map $\mathbb{R}^d \mapsto \mathbb{R}^m$, $A = [\mathbf{a}_1, \cdots, \mathbf{a}_r] \in \mathbb{R}^{m \times r}$. Therefore, the new objective function $\mathcal{Q}$ that offers a clearer connection and easier access to the model parameter A is

$$\min_{A \in \mathbb{R}^{m \times r}} \mathcal{Q}(A) = \left\| \frac{1}{r} sgn(\bar{K}_l A)(sgn(\bar{K}_l A))^T - S \right\|_F^2 \quad (4)$$

Optimization: Since the objective function $\mathcal{Q}$ is neither convex nor smooth, two optimization schemes are employed: 1) Spectral Relaxation [15] is applied to drop the sign functions and hence convexifies the object function; 2) Sigmoid Smoothing is employed to replace $sgn()$ with the sigmoid-shaped function $\varphi(x) = 2/(1 + exp(-x)) - 1$. Consequently, the objective function $\mathcal{Q}$ is able to be minimized using the standard gradient descent technique. After obtaining the discriminative hash functions via optimizing $\mathcal{Q}$, high-dimensional SIFT image features can be mapped into informative binary bits which are further indexed into a hash table for real-time search of similar histopathological images.

3. EXPERIMENTS

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

Experimental Setting:

2646 images (around 2250K pixels) are sampled from 657 larger region-of-interests images (e.g. 5K×7K) of breast microscopic tissue, which are gathered from 116 patients[1], labeled as the benign category (UDH) and the actionable category (ADH and DCIS). 25% of these patients in each category are randomly selected as the testing set and the other cases are used for training. All the experiments are conducted on a 3.40GHz CPU with 4 cores and 16G RAM, in a MATLAB implementation.

Around 1500 to 2000 SIFT descriptors are extracted from each image and quantized into sets of cluster centers using bag-of-words, in which the feature dimension equals the number of clusters. This hashing-based method is evaluated on two tasks: image classification (i.e., benign v.s. actionable category) and image retrieval. The classification is achieved using the majority logic of top retrieved images. In the classification task, we compare with the classical classifiers such as support vector machine (SVM) and AdaBoost, k-nearest neighbors (kNN), which have been used for histopathological image analysis [3, 5, 6, 16]. All kernel selections and parameters are optimized by cross-validation. In addition, we also compare with kNN after applying principal component analysis (PCA) as a dimension reduction method. In the evaluation of image retrieval, we just compare with kNN (with and without PCA), since SVM and Adaboost are not normally applicable to this task.

[1]Data is provided by the Clarion Pathology Lab, Indianapolis and the Computer and Information Science Department, IUPUI, Indiana, using ScanScope digitizer at 40× magnification.

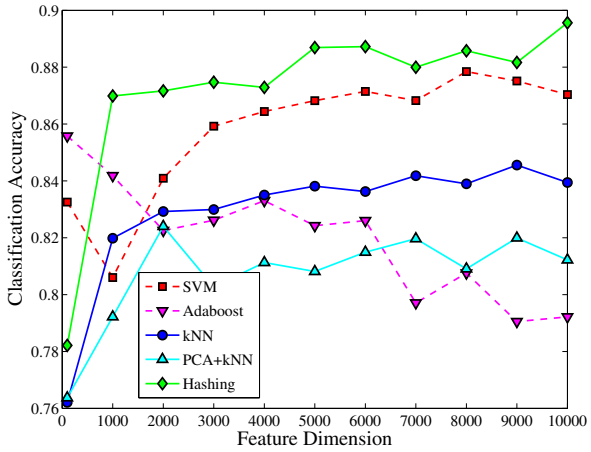

Fig. 2. Comparison of the classification accuracy with different dimensions of features.

Evaluation of Image Classification: All methods are evaluated on different dimensions of SIFT quantization, ranging from 100 to 10000. We use hashing method to compress all features to 48 bits (only 6 bytes). For fair comparison, we also use PCA to compress all features to 48 dimensions. Note that PCA results are float numbers (4 to 8 bytes for each float), which are much larger than hashing results, so such comparison actually favors the dimension reduction method.

Fig. 2 shows the comparisons of the classification task. Most methods achieve better accuracy with higher dimensional features. This is very intuitive as finer quantization of SIFT features usually provides richer information. Particularly, since the SIFT interest points cover most nuclei regions in images, fine quantization (i.e., high-dimensional feature) indicates the analysis on small scales. One exception is that the accuracy of Adaboost drops when increasing the feature dimensions. The reason is that Adaboost is essentially a feature selection method, which only choose an effective subset of features for the classification. Therefore, it may lose important information, especially in high dimensional space. kNN-based classification also achieves good accuracy. After applying PCA-based dimension reduction, its accuracy is usually lower than using the original features, due to the information loss in compression. Our hashing method and SVM are generally better than kNN, owing to the supervised information (i.e., labels of similar and dissimilar pairs in hashing). Note that our hashing method only needs a small amount of supervision, in this case, similar or dissimilar pairs of 300 images. It compares favorable to all other methods when the feature dimension is larger than 1000. The overall classification accuracy is 89.6% (90.5% for benign and 87.6% for actionable category) when using 10000 dimensional features. It is 3% to 11% better than other methods.

	kNN	kNN+PCA	Hashing
P@10	0.809	0.798	0.877
P@20	0.794	0.792	0.876
P@30	0.786	0.785	0.876
Time(s)	7.56	0.07	<0.01
Memory	133.59Mb	0.64Mb	10.26Kb

Table 1. Comparison of retrieval precision at top-10, 20 and 30 results, along with the memory cost of training data and query time of all testing images.

Evaluation of Image Retrieval: We have also conducted experiments on image retrieval using 10000 dimensional features. The retrieval precision is reported in Table 1, along with the query time and memory cost. The results are quite consistent with the image classification. The precision of hashing method is nearly 88% (87.5% for benign and 87.9% for actionable category). This is significantly better than kNN and kNN with PCA, i.e., around 10% margin. In addition, the memory cost and runtime is also considerably reduced. Therefore, this method is more applicable to large scale databases (e.g., millions of images) than other methods. Fig. 3 shows four examples of our image retrieval results. The local differences of certain images are very subtle. Our accurate results demonstrate the efficacy of the proposed method and the feature, which captures local texture and appearance. These retrieved images are clinically relevant and thus very useful for decision support.

4. CONCLUSION

In this paper, we introduced a *scalable image retrieval framework* for histopathological image analysis. Specifically,we focused on hashing-based retrieval methods, and investigated a 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 the future, we will examine other types of features, especially features stemming from segmentation and architectures. In addition, we will incorporate feature fusion techniques into hashing methods. Therefore, multiple types of features can be combined to improve the retrieval accuracy. We will also evaluate our framework on more applications in histopathological image analysis.

5. REFERENCES

[1] J. C. Caicedo, A. Cruz, and F. A. Gonzalez. Histopathology image classification using bag of features and kernel functions. In *Artificial Intelligence in Medicine*, pages 126–135. 2009.

[2] M. Datar, N. Immorlica, P. Indyk, and V. S. Mirrokni. Locality-sensitive hashing scheme based on p-stable distributions. In *SoCG*, pages 253–262. ACM, 2004.

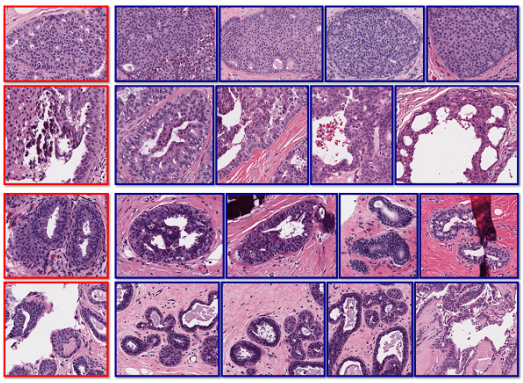

Fig. 3. Four examples of our image retrieval (query marked in red, and retrieved images marked in blue). The first two rows are actionable, and the last two rows are benign.

[3] 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 *ISBI 2008*, pages 496–499, 2008.

[4] M. Dundar, S. Badve, G. Bilgin, V. Raykar, R. Jain, O. Sertel, and M. Gurcan. Computerized classification of intraductal breast lesions using histopathological images. *TBME*, 58(7):1977–1984, 2011.

[5] D. J. Foran, L. Yang, et al. Imageminer: a software system for comparative analysis of tissue microarrays using content-based image retrieval, high-performance computing, and grid technology. *JAMIA*, 18(4):403–415, 2011.

[6] M. N. Gurcan, L. E. Boucheron, A. Can, A. Madabhushi, N. M. Rajpoot, and B. Yener. Histopathological image analysis: A review. *IEEE R-BME*, 2:147–171, 2009.

[7] B. Kulis and K. Grauman. Kernelized locality-sensitive hashing for scalable image search. In *CVPR*, 2009.

[8] W. Liu, J. Wang, R. Ji, Y.-G. Jiang, and S.-F. Chang. Supervised hashing with kernels. In *CVPR*, pages 2074–2081, 2012.

[9] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, Nov. 2004.

[10] H. Müller, N. Michoux, D. Bandon, and A. Geissbuhler. A review of content-based image retrieval systems in medical applicationsłclinical benefits and future directions. *IJMI*, 73(1):1–23, 2004.

[11] 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 Medical Imaging*, 6(1):14, 2006.

[12] S. Sanati and D. C. Allred. *Pre-Invasive Disease: Pathogenesis and Clinical Management*, chapter 5 The Progression of Pre-invasive to Invasive Cancer. Springer New York, 2011.

[13] R. Siegel, D. Naishadham, and A. Jemal. Cancer statistics, 2013. *CA: a cancer journal for clinicians*, 63(1):11–30, 2013.

[14] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *ICCV*, 2003.

[15] Y. Weiss, A. Torralba, and R. Fergus. Spectral hashing. In *NIPS*, pages 1753–1760, 2008.

[16] 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. *TITB*, 13(4):636–644, 2009.