Size-scalable content-based histopathological image retrieval from database that consists of WSIs

Yushan Zheng, Zhiguo Jiang*, *Member, IEEE,* Haopeng Zhang**, *Member, IEEE,* Fengying Xie, Yibing Ma, Huaqiang Shi and Yu Zhao

***Abstract*—Content-based image retrieval (CBIR) has been widely researched for histopathological images. It is challenging to retrieve contently similar regions from histopathological whole slide images (WSIs) for regions of interest (ROIs) in different size. In this paper, we propose a novel CBIR framework for database that consists of WSIs and size-scalable query ROIs. Each WSI in the database is encoded into a matrix of binary codes. When retrieving, a group of region proposals that have similar size with the query ROI are firstly located in the database through an efficient table-lookup approach. Then, these regions are ranked by a designed multi-binary-code-based similarity measurement. Finally, the top relevant regions and their locations in the WSIs as well as the corresponding diagnostic information are returned to assist pathologists. The effectiveness of the proposed framework is evaluated on a fine-annotated WSI database of epithelial breast tumors. The experimental results have proved that the proposed framework is effective for retrieval from database that consists of WSIs. Specifically, for query ROIs of 4096×4096 pixels, the retrieval precision of the top 20 return has reached 96% and the retrieval time is less than 1.5 second.**

***Index Terms*—histopathological image, CBIR, WSI, binary code, hashing, breast cancer**

I. Introduction

The diagnosis of cancer using histopathological images is a challenging task. Specifically for breast cancer, the precise diagnosis with histopathological images is a difficult work due to the diversity of breast lesions [1] and the subtle difference between sub-categories of lesions in histopathological images. In this situation, content-based histopathological image retrieval (CBHIR) is developed based on histopathological image analysis (HIA) approaches [2], [3], [4] to aid pathologists. For a query image, CBHIR can search for the database and return images that are contently similar to the query image. Using diagnostic information of these similar cases for reference, doctors can comprehensively understand the case and reach a more reasonable diagnosis.

The research of CBHIR for histopathological images can date back to 1998. In [5], [6], Comaniciu et al. introduced a CBIR framework to retrieve the images that contain similar cell nuclei for a query image. Then, some researchers [7], [8], [9], [10] utilized the classical image features to depict the histopathological images and achieved the retrieval for histopathological images in cellar level and sub-image level. However, histopathological images are much more complex than natural images that using low-level features cannot discriminatively represent the histopathological images. In recent years, researchers applied the high-level feature extraction models, such as manifold learning [11], [12], semantic analysis [13], [14], [15], [16], [17], [18], spectral embedding [19], etc., to the histopathological image representation, which improved the performance of CBHIR. The approaches mentioned above have been concentrated on the feature extraction from histopathological images and the similarities among images are measured based on feature vectors, which will raise a high computation when retrieving from a large-scale database.

To satisfy the application for database consisting of massive histopathological images, the efficiency of CBHIR is considered in the recent researches [20]. Zhang et al. [21][22] introduced a supervised hashing method [23] into the CBHIR. Instead of using high dimensional feature vectors to represent histopathological images, they encoded each image into an array of binary codes and stored it within tens of bits. Then the similarities among images can be measured by Hamming distance, which is able to be calculated very efficiently using bitwise operations by computer.

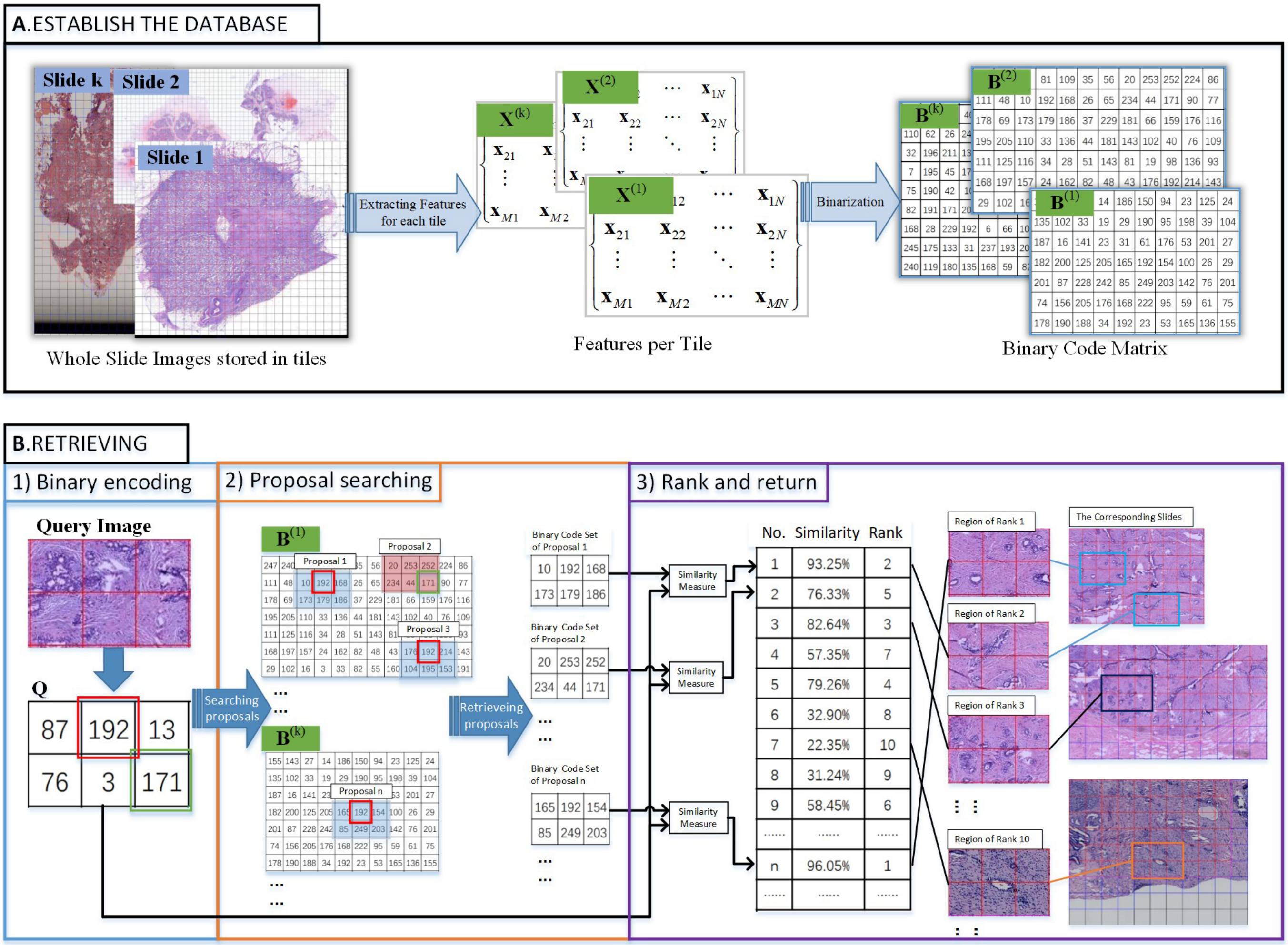
Zhang et al. [21], [22] have provide an efficient CBHIR approach for database of individual images. However, the practical digital slide is stored in a spatially continuous image with a size of more than $10K \times 10K$ pixels and the region of interest (ROI) is size-scalable according to different diagnostic conditions. This makes it a challenging task to retrieve eligible regions from database consisting of whole slide images (WSIs). More recently, Ma el al. [24] proposed a binary histopathological representation based on a latent Dirichlet allocation (LDA) model and applied the retrieval framework to WSIs following a sliding window (SW) paradigm. It provided a preliminary approach for CBHIR from WSIs. Nevertheless, three issues are required to tackle when applied CBHIR to a practical WSI database. First, to achieve a precise retrieval from WSIs, a group of regions need to be sampled in overlapping manner throughout the WSIs. When retrieving, the query ROI needs to be compared with all the regions in the large database, which causes a high computation. Second, the size of the query ROI varies greatly according to the diagnostic requirement. To satisfy the various size of query

This work was supported by the National Natural Science Foundation of China (No. 61371134, 61471016, and 61501009) and project of Motic-BUAA Image Technology Research and Development Center. *Asterisk indicates corresponding author.*

Yushan Zheng, Zhiguo Jiang*, Haopeng Zhang**, Fengying Xie and Yibing Ma are with Image Processing Center, School of Astronautics, Beihang University, Beijing, 100191, China. (e-mail: yszheng, jiangzg, zhanghaopeng, xfy_73, mayibing@buaa.edu.cn)

Huaqiang Shi and Yu Zhao are with Motic (Xiamen) Medical Diagnostic Systems Co. Ltd., Xiamen 361101, China. (e-mail: shihq, zhaoy@motic.com)

Huaqiang Shi is chief physician, department of pathology, Peoples Liberation Army Air Force General Hospital, Beijing 100142, China.



87	192	13
76	3	171

No.	Similarity	Rank
1	93.25%	2
2	76.33%	5
3	82.64%	3
4	57.35%	7
5	79.26%	4
6	32.90%	8
7	22.35%	10
8	31.24%	9
9	58.45%	6
......	......	......
n	96.05%	1
......	......	......

Fig. 1. Flowchart of the proposed framework, where the binary code is displayed in decimal numeral for clearness,

ROIs, multiple retrieval models and the corresponding retrieval databases need to be established. Third, when the query ROI has a large size, e.g. $4,000 \times 4,000$ pixels, quantifying the features in the image into single binary code will weaken the local patterns of the histopathological image, thus decrease the precision of retrieval.

In this paper, we propose a novel content-based histopathological image retrieval framework for a database consisting of WSIs, in which all the three issues stated above are simultaneously considered and resolved. Instead of modeling a WSI in the retrieval database using individual images, we propose to encode the entire WSI into a matrix of binary codes. When retrieving, a group of region proposals that have the similar size of the query ROI are located in the binary matrices via looking-up a pre-established hashing table and then are ranked by a designed multi-binary-code-based similarity measurement. Using our framework, the retrieval with ROIs in different size can be completed using one model and within one database, which sharply reduces the computation and storage compared to the present retrieval framework. The proposed framework was evaluated on a database that contains epithelial breast tumors. The experiments have certified the effectiveness of our method. A preliminary version of this work has been reported [25]. In this paper, we provide more details of methodology, present further evaluations to demonstrate the capability of our method, and compare our method with the state-of-the-art breast image retrieval frameworks.

The remainder of this paper is organized as follows. Section II introduces the proposed retrieval framework. The experiment is presented in Section III. Finally, Section IV summarizes the present contributions and suggests directions for future work.

II. Method

The flowchart of the proposed framework is illustrated in Figure 1. A retrieval database is first established by encoding WSIs into matrices of binary codes. Then, the retrieval can be achieved through 3 steps: 1) binary encoding, 2) proposal searching, and 3) ranking & returning. Next, we introduce the technical details of our retrieval framework.

A. The binariztion of WSIs in the database

The database used in our methods is established with binary codes matrices of WSIs. As shown in Figure 1A, a WSI is first divided into non-overlapping square tiles. Let T_{ij} denotes the tile in the i-th row and the j-th column in the WSI. Then, the feature extraction of the tile can be represented as:

$$\mathbf{x}_{ij} = f(T_{ij}), \tag{1}$$

where, $\mathbf{x}_{ij}$ is a feature vector and $f(\cdot)$ denotes a feature extraction method. Next, the binary code b_{ij} for Tile T_{ij} can be calculated by

$$b_{ij} = \mathbf{h}(\mathbf{x}_{ij}), \tag{2}$$

where $\mathbf{h}(\cdot) = \{h_1(\cdot), h_2(\cdot), \ldots, h_K(\cdot)\}$ denotes a set of binary functions. K is the function number, namely the bit number of the binary code b_{ij}.

Considering the practical situation of digital pathology platforms (e.g. MoticGallery[1]) where more than 10^5 digital WSIs of diagnosed cases are collected but very few of them are precisely labeled by pathologists, we propose to establish the framework with unsupervised feature extraction methods and binary methods.

Repeating the operation of Eq. 1 and Eq. 2 throughout all the tiles in the WSI, a matrix that consists of binary codes can be obtained and represented as

$$\mathbf{B} = \begin{pmatrix} b_{11} & b_{12} & \cdots & b_{1N} \\ b_{21} & b_{22} & \cdots & b_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ b_{M1} & b_{M2} & \cdots & b_{MN} \end{pmatrix}, \tag{3}$$

where M and N denote the row and column number of tiles in the WSI. Notice that any submatrix[2] generated with contiguous rows and columns of matrix $\mathbf{B}$ can be regarded as a description of the corresponding region in this WSI. Therefore, the regions that are eligible to be retrieved from the WSI are determined by all the feasible submatrices of $\mathbf{B}$.

Let $\mathbf{B}_l$ denote the binary matrix of the l-th WSI. Then a database consists of L WSIs is defined using a set

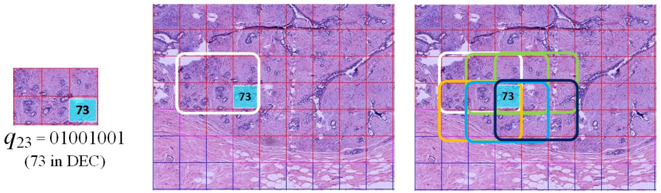
$$\mathscr{D} = \{\mathbf{B}_1, \mathbf{B}_2, \ldots, \mathbf{B}_L\}.$$

Suppose the bit number $K = 64$. Then a binary code b_{ij} can be stored with 8 bytes in memory. Setting the tile size as 512×512 pixels, the storage of the code matrix for a WSI with $50K \times 50K$ pixels (a medium size under a $20\times$ lens for breast tumor slide) is less than $80K$ bytes.

B. Retrieval with binary codes

Based on the database $\mathscr{D}$, a retrieval approach for size-scalable query ROI is designed in this paper, which can be divided into three steps (as shown in Figure 1(B)). The query ROI is first encoded into a set of binary codes. Then, a group of regions that have the similar size to the query ROI are proposed from the database $\mathscr{D}$. After ranking the similarities between the query ROI and these proposed regions, the most relevant regions as well as their locations in the WSIs are finally determined and returned to doctors.

The details of the three steps are presented as follows:



(a) Query Roi (b) Proposal regions defined by C_1 (c) Proposal regions defined by C_2

Fig. 2. An instance of proposal regions determined by a tile, where (a) is the query image, in which the binary code q_{23} = 01001001 (73 in DEC), (b) marks the region proposal defined by C_1, and (c) shows the region proposals defined by C_2.

1) Binary Encoding: The binarization of a query ROI is similar to a WSI of the database. The ROI is first divided into square tiles that have the same size of those in the database. Then, these tiles are encoded into a binary matrix $\mathbf{Q}$ using Eq. 1 and 2. In practical applications, the query ROI is cropped in terms of containing as much as tiles for encoding, since regions that can be retrieved from $\mathscr{D}$ are mosaicked with tiles. Then, the retrieval for size-scalable query image is designed based on the binary codes of these tiles.

2) Proposal Searching: To refine the searching scope and reduce the computation in similarity measuring, a set of regions are previously proposed from all the feasible returned regions. Intuitively, the scale of the returned regions should have the similar size to the query one, thereby the returned regions are restricted to submatrices that have the same size with $\mathbf{Q}$. Nevertheless, the number of feasible submatrices is even so large that it is obviously unreasonable to consider all the size-feasible submatrices in $\mathscr{D}$, especially for a large-scale database.

Notice the property of binary encoding that samples sharing an equal binary code should be more similar than those with different codes. Basing on this property, we propose to locate tiles that share the same binary code with tiles of the query ROI in the database. Then, the submatrices including these tiles are extracted from $\mathscr{D}$ as the region proposals, as illustrated in Figure 1B(2). This process can be accomplished very efficiently via table lookup operation with a pre-established hashing table. Let $sub(B_l)$ denote the collection that contains all the submatrices included in B_l. Then, two types of region proposals are defined by collections:

$$\begin{aligned} C_1 : \mathscr{P} = \{&\mathbf{P} | \exists q_{mn} \in \mathbf{Q}, \exists p_{mn} \in \mathbf{P}, \mathbf{P} \in sub(B_l), \\ &B_l \in \mathscr{D}, s.t. q_{mn} = p_{mn}\}, \end{aligned} \tag{4}$$

$$\begin{aligned} C_2 : \mathscr{P} = \{&\mathbf{P} | \exists q_{mn} \in \mathbf{Q}, \exists p_{m'n'} \in \mathbf{P}, \mathbf{P} \in sub(B_l), \\ &B_l \in \mathscr{D}, s.t. q_{mn} = p_{m'n'}\}, \end{aligned} \tag{5}$$

where $\mathbf{P}$ denotes a submatrix that has the same size with $\mathbf{Q}$, p_{mn} and q_{mn} denote the binary code in the m-th row and the n-th column of matrix $\mathbf{P}$ and $\mathbf{Q}$, respectively. As illustrated in Figure 2, C_1 is a tile-location-associated scheme while C_2 is a densely sampling scheme for the selection of proposals. Apparently, the scale of $\mathscr{P}$ is positively correlated with the scale of $\mathscr{D}$. Besides, it is negatively correlated with K (the bin number of the binary code), since the tiles will be described

[1] http://med.motic.com/MoticGallery/[accessible 2017-06-12]

[2] Submatrix is referred specifically to the submatrix generated with contiguous rows and columns of a matrix in this paper.

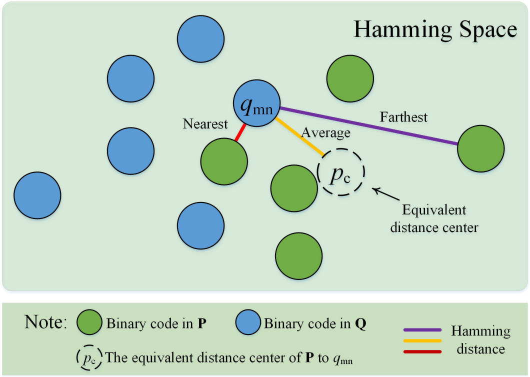


Fig. 3. An instance for the three definitions of distance in hamming space, where p_c is a virtual center to describe the average distance between q_{mn} and all the elements in $\mathbf{P}$.

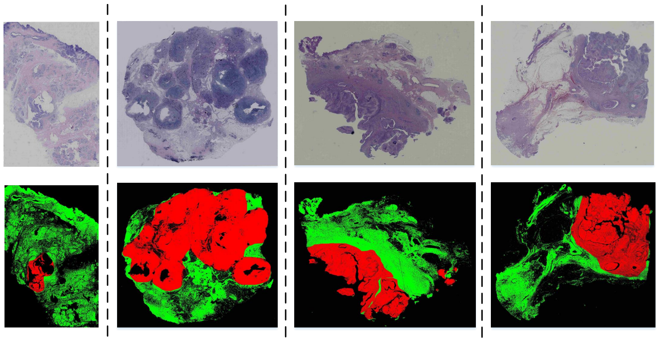


Fig. 4. Instances of complete annotation, where all the regions of epithelial tumors in a WSI are annotated and displayed in red, non-epithelial/normal regions are displayed in green, and adipose tissue and background are displayed in black .

more discriminatively as the binary codes become longer and the tiles that sharing the same code will reduce.

3) Ranking & returning: The final retrieval results are selected from $\mathscr{P}$ by ranking the similarities between the query code matrix $\mathbf{Q}$ and the proposals in $\mathscr{P}$. For the query ROI and the proposals are represented by multiple binary codes (MBC), a simple Hamming distance cannot measure the similarities among them. Therefore, MBC-based distances are defined in this paper. Considering a certain binary code q_{mn} in the query matrix $\mathbf{Q}$, the distance from q_{mn} to a proposal matrix $\mathbf{P}$ is defined in Hamming space according to three basic principles: the average, nearest, and farthest distance. As illustrated in Figure 3, the "Nearest" and "Farthest" distance from q_{mn} to $\mathbf{P}$ are defined as the distance between q_{mn} and the nearest and farthest elements in $\mathbf{P}$, respectively. And the mean distance between q_{mn} and all the elements of $\mathbf{P}$ is defined as the "Average" distance. Based on the definitions, three types of distances from matrix $\mathbf{Q}$ to $\mathbf{P}$ are defined:

$$D_{Ave}(\mathbf{Q},\mathbf{P}) = \frac{1}{b}\sum_{q_{mn}\in\mathbf{Q}}\left(\frac{1}{b}\sum_{p_{ij}\in\mathbf{P}} d_h(q_{mn},p_{ij})\right), \tag{6}$$

$$D_{Near}(\mathbf{Q},\mathbf{P}) = \frac{1}{b}\sum_{q_{mn}\in\mathbf{Q}}\min_{p_{ij}\in\mathbf{P}} d_h(q_{mn},p_{ij}), \tag{7}$$

$$D_{Far}(\mathbf{Q},\mathbf{P}) = \frac{1}{b}\sum_{q_{mn}\in\mathbf{Q}}\max_{p_{ij}\in\mathbf{P}} d_h(q_{mn},p_{ij}), \tag{8}$$

where b denotes the number of binary codes in matrix $\mathbf{P}$ and $\mathbf{Q}$ (i.e. the number of tiles included in the query ROI), and $d_h(p,q)$ represents the hamming distance between the binary code p and q, namely the number of different bins in comparison of p and q. As defined, the more similar Q and P, the closer the three distances should be. These three distances perform differently when they are used for similarity measuring. Eq. 6 describes a generally similarity between the tiles in the query ROI and those in a proposed region. While Eq. 7 is defined to character the most similar tile in $\mathbf{P}$ to each tile in $\mathbf{Q}$, which encourages the model to return regions containing the most similar local appearance to the query ROI. In contrast, Eq. 8 is sensitive to the most dissimilar tile in $\mathbf{P}$ to that in $\mathbf{Q}$, which impels the model to return regions that have less distinct local appearance with the query ROI. Notice that the distance $d_{Near}(\mathbf{Q},\mathbf{P})$ ($d_{Far}(\mathbf{Q},\mathbf{P})$) is not equivalent to $d_{Near}(\mathbf{P},\mathbf{Q})$ ($d_{Far}(\mathbf{P},\mathbf{Q})$). To describe an interactively similar relationship between $\mathbf{P}$ and $\mathbf{Q}$, two further distances are defined:

$$D_{inter-Near}(\mathbf{Q},\mathbf{P}) = d_{Near}(\mathbf{Q},\mathbf{P}) + d_{Near}(\mathbf{P},\mathbf{Q}), \tag{9}$$

$$D_{inter-Far}(\mathbf{Q},\mathbf{P}) = d_{Far}(\mathbf{Q},\mathbf{P}) + d_{Far}(\mathbf{P},\mathbf{Q}). \tag{10}$$

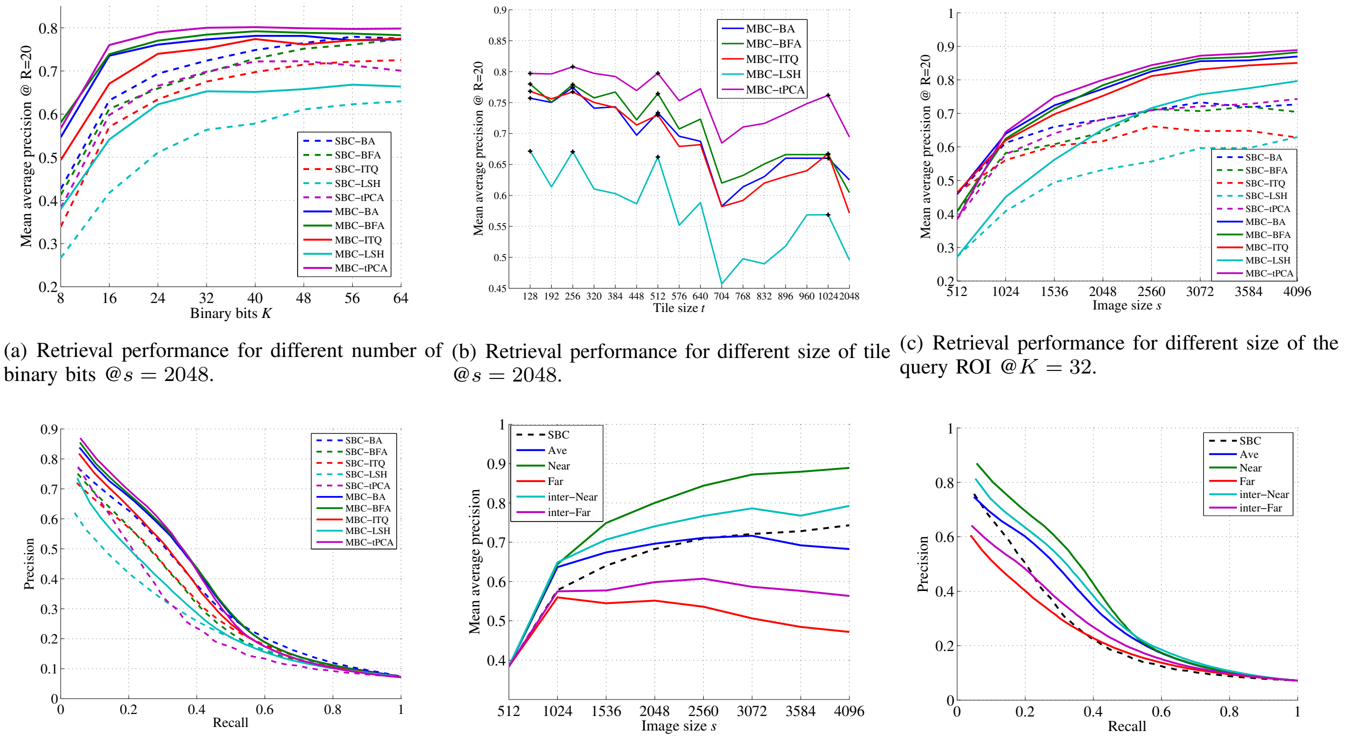
Choosing one of the five distances as the similarity measurement, the proposed regions indicated by the code matrices in $\mathscr{P}$ are ranked in terms of similarity. The lower the value obtained by Eq. 6-10, the more similar the two regions. Finally, the top-similar regions with their locations in the corresponding WSIs are returned as the retrieval result. The effectiveness of the five designed distances are discussed in the experiment section.

III. Experiment

A. *Experimental Data*

The experimental images used in this research are supplied by Motic[3] (Motic database). There are 145 HE-stained whole slide images (WSIs) including epithelial breast tumors under 20x magnification (the spatial resolution is 1.2 μm/pixel). The 145 WSIs are annotated by the pathologists from the air force general hospital of China, of which 50 WSIs are completely annotated and the other 95 WSIs are locally annotated. These annotations are labeled as 15 sub-categories of epithelial breast tumors according to the world health organization (WHO) standard [1]. Figure 4 gives 4 instances of complete annotation, where the regions of epithelial breast tumors are illustrated in red, regions of other tissue are illustrated in green, and the adipose tissue and the background regions are displayed in black.

[3] Motic (Xiamen) Medical Diagnostic Systems Co. Ltd., Xiamen 361101, China



(a) Retrieval performance for different number of binary bits $@s = 2048$.

(b) Retrieval performance for different size of tile $@s = 2048$.

(c) Retrieval performance for different size of the query ROI $@K = 32$.

(d) Precision-Recall curves for different hashing methods $@s = 2048$.

(e) Retrieval performance for different different similarity measurement.

(f) Precision-Recall curves for different similarity measurement $@s = 2048$.

Fig. 5. Retrieval performance of the multi-binary-code-based similarity measurement on database that consists of individual images.

B. Experimental setting

We first conducted experiment to validate the multi-binary-code-based similarity measurement using individual images and then conducted experiments to evaluate the effectiveness of the proposal selection approach on WSIs database. Finally, four state-of-the-art retrieval frameworks were compared. In the experiment, the bag of features (BoF) based on SIFT descriptors are used to represent histopathological images, of which the effectiveness for CBHIR is certified in [14], [21]. To evaluate the universality of our framework, 5 unsupervised binarization methods are validated. They are Locality-Sensitive Hashing (LSH) [26], thresholded PCA (tPCA)[27], Iterative Quantization (ITQ) [28], binary factor analysis (BFA) [29] and Binary Autoencoders (BA) [29].

In application of CBHIR, the retrieved regions that include the same sub-category of the query ROI are desired by doctors. Therefore, in the experiment, the returned images and regions including the same type of epithelial tumor with the query ROI are regarded as the correct retrieval results. And for query ROIs of non-tumor/normal tissues, the results containing no epithelial tumor area are regarded as correct. The mean average precision (MAP) of the top 20 returned images for all the testing images and the precision-recall (PR) curve are used as the assessment metrics. All the methods are implemented using MATLAB 2013a on a PC with 12 cores of 2.10 GHz.

C. Effectiveness of the proposed similarity measurement

Our framework is a size-scalable image retrieval method. The regions of interest (ROIs) in different size are sampled to evaluate the retrieval performance of the proposed similarity measurement. Specifically, the size of ROIs is ranged from 512×512 to 4096×4096 pixels with an image side step of 512 pixel. For each size, 6600 images are randomly sampled from the annotated regions of the 95 locally annotated WSIs for evaluation. The retrieval performance of each query size is validated independently through a 5-fold cross-validation [30], in which the one fifth images of each fold are used as the query images and the others are regarded as the database. To present the advantage of our method for large query images, the performance of retrieval using single binary-code (used in [24]) is also evaluated. For clearness, the results of the proposed MBC-based methods are prefixed by 'MBC-' and the results using the single binary code (SBC) are prefixed by 'SBC-'.

1) Overall performance: Figure 5 presents the retrieval performance of different binarization methods. It can be seen that the retrieval precisions using MBC-based methods are generally greater than those measured using SBC-methods. SBC-methods encoded the entire image as one binary code, which weakened the local information of the histopathological image. While, MBC-methods divided an image into tiles and represented the image using multiple binary codes, by which more details of the histopathological images are preserved. Therefore, the MBC-methods achieved a better performance.

2) The bits of binary code K: For hash-based retrieval framework, the bit number of the binary code K is one of the most important parameter, which directly influences the accuracy of retrieval and the storage space of the database. An efficient retrieval framework shall utilize as few as bits of code to maintain a high retrieval accuracy. Setting the ROI size $s = 2048$ and D_{Near} as the similarity measurement, the sensitivity of K is evaluated. Figure 5(a) presents the MAPs obtained under different K for the 10 considered models. In general, the retrieval precision is positively correlated with K. And the performance of MBC-based methods is more robust to K than SBC-based methods, since the MAPs of the MBC-based methods raise to stable values much faster than the SBC-based methods as K enlarges. Specifically for MBC-tPCA method, $K = 32$ is enough when applied to this database.

3) The size of tile t: Another important parameter for MBC-based similarity measurement is the side length of tile (t). Figure 5(b) gives the retrieval performance with different t. It is noting that the local optimal points (marked by "+" on the figure) are appeared when the ROI size s is divisible by t, and the accuracy declined when s was indivisible by t. The reason is that the amount of information in the query ROI is reduced by the cropping operation when s is indivisible by t. In the local optimal points, $t = 128, 256, 512$ achieves relatively equivalent performance for tPCA method. But, more tiles will cause more computation in the proposed measurements (Eq.6-9). Therefore, t is finally set as 512.

4) The binarization methods: As presented in Figure 5(c,d), the MBC-tPCA achieved the best performance. The binarization model tPCA is designed based on the principle component analysis (PCA), which is a parameter-free model. The performance of tPCA is only influenced by the bits of binary code K. In the evaluated binarization models, BA also achieved competitive performance in this experiment. Especially, BA obtained the highest retrieval accuracy in the five SBC-based methods. However, the modeling of BA follows an alternating optimization over two steps [29] and several parameters in BA need to be adjusted for optimal performance, which is complicated. For consideration of complexity and robustness, tPCA is finally determined as the hashing model of the proposed framework.

5) Performance for different size of query ROIs: Setting $K = 32$ and $t = 512$, the retrieval performance of different sizes of query images are illustrated in Figure 5(c). Obviously, using MBC-methods, the performance of retrieval obtained a great improvement under each query size (except the size of 512×512, for the ROIs containing only one tile that the MBC-based methods are equivalent to the SBC-based methods). In addition, the advantage of the MBC-based methods increases as s growing. It is because that a large query image contains more local information than the small one. Using MBC-based approaches to describe the large image maintains more details about tumors, thus performs better than the SBC-based methods.

6) Performance for different returned numbers: By adjusting the number of the returned images with a step of 10, a precision-recall (PR) curve can be obtained. The PR curve corresponding to $K = 32$ and $s = 2048$ is drawn in Figure 5 (d), which shows the general advantage of the proposed MBC-CBIR framework. Specifically, the difference between MBC-tPCA and SBC-tPCA is remarkable.

TABLE I
RETRIEVAL PERFORMANCE ON WSI DATABASE @$K = 48$.

Image size s		1024	2048	3072	4096
MAP	C_1	0.880	0.921	0.925	0.937
	C_2	0.959	0.964	0.964	0.961

TABLE II
AVERAGE NUMBER OF REGION PROPOSALS FOR WSIS USING DIFFERENT PROPOSAL SELECTION APPROACHES.

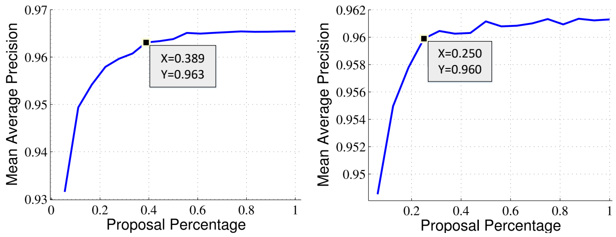
Image size s		1024	2048	3072	4096
Tile number t		4	16	36	64
#Proposal	SW[24]	480K	461K	441K	425K
	C_1	15.3	23.2	46.2	58.4
	C_2	61.2	242.2	1663.3	3736.2

7) Performance for different similarity measurements: For the five similarity measurements defined in section II-B3, the experimental results are illustrated in Figure 5(e,f), where the binarization method is chosen as tPCA. Overall, the Nearest-distance achieved the best performance. In the five measurements, only the two distances designed following the nearest principle achieved better performance than the SBC-based method for all the sizes of testing images. And, the other distances did not perform well. More seriously, the MAPs for distances defined by the farthest principle decreased when the query ROI enlarged. Referring to the definition, the three distances are designed to character more about global similarity, for which the local patterns of the image are weakened, especially for large ROIs. While, the Nearest and inter-Nearest distances pay more attention to the similarities in local patterns, which is more significant for diagnosis. Therefore, the two measurements maintain a high retrieval precision for large query ROIs. Consequently, the Nearest and inter-Nearest distances are more appropriate for the proposed CBHIR framework.

D. Effectiveness of the proposal selection approach

When applying MBC-based retrieval to WSIs, the proposal selection approach introduced in section II-B2 is required. Choosing tPCA as the binarization method and Nearest-distance (Eq. 7) as the similarity measurement, we conducted experiments to evaluate the effectiveness of the proposal selection approaches on WSI database. The 50 completely annotated WSIs were regarded as the database. And for each testing size, the 6600 images sampled from the annotated regions of the locally annotated WSIs were used as the query ROIs. Similar to the evaluation of individual images, the MAP for top 20 returned regions are used as metric.

1) Overall Performance: Table I reports the retrieval results using the two region proposal methods for different size of query images. Overall, the proposed framework achieves considerable retrieval precisions. Specifically, using C_2-type (Eq.5) region proposals, the MAP is over 95% for the four



(a) Image size s=3072

(b) Image size s=4096

Fig. 6. Retrieval performance using different percentage of region proposals for large query ROIs.

query sizes, which are better than C_1-type region proposals. Notice that dense regions around a certain tile in WSIs were proposed by C_2, which would bring redundancy to the retrieval result. Therefore, only the most similar region around a tile was preserved and the redundant results around the same tile were excluded in the ranking procedure.

2) Leveraging the two region proposing methods: The number of regions using C_1 and C_2 proposal selection schemes are reported in Table II, in which the estimation for the sliding window (SW) paradigm used in [24] are also given for comparison. It shows that the two region proposing methods sharply reduced the scope of searching. In the experiment, C_2 achieved a better performance but yielded a large amount of region proposals for large query ROIs. To find a balance of the two methods, we reduced the C_2-type region proposals for 3072×3072 and 4096×4096 query ROIs through randomly sampling. Figure 6 presents the MAPs that vary with the percentage of region proposals used in retrieval. It can be seen that the MAPs are over 96.0% when the percentage of proposal increases to 38% and 25% for 3072×3072 and 4096×4096 query images, respectively. And using more region proposals, the MAP increased little. Therefore, 38% and 25% proposals are enough for retrieval with 3072×3072 and 4096×4096 query ROIs.

3) Effect of the binary bits: The performance of proposal searching is also influenced by K. Table III presents retrieval performance for different K. When K is too small (e.g. $K = 12$), the tiles in WSI cannot be discriminatively represented and too many tiles are assigned to the same binary code, which will result in a high computation in the ranking & returning step. When $K > 48$, both the proposal number and MAP change little, which indicates that the binarizaition encoding is already redundant for the representation of the 50 WSIs. Considering both the precision and time consumption of retrieval, $K = 48$ are the most appropriate for the proposed framework. The time cost for the retrieval stage is also reported in Figure III. The average feature extracting times for image size $s =$1024, 2048, 3072, and 4096 are 0.11, 0.36, 0.78, and 1.30, respectively. Therefore, a retrieval for a query ROI with a size of under 4096×4096 pixels can be completed within 1.5 second. Generally, choosing an appropriate number of binary bits according to the amount of WSIs in the dataset can effectively control the computational efficiency and retrieval performance, achiving a considerable retrieval precision and a relatively short time for large-scale WSI database.

E. Visual results of the proposed model

Figure 7 visualizes two instances of retrieval for the same ROI in different scale and Figure 8 illustrates the results retrieved by different similarity measurements for the two query ROIs, which shows that the retrieval performance is consistent with the numerical assessment. In Figure 8, the top-ranked regions returned by the Nearest distance are the most relevant to the query ROI and the regions returned by the inter-Nearest-distance relatively cover more diagnosed cases. Therefore, we recommend using the two similarity measurements in practical applications.

F. Comparison with other retrieval frameworks

To thoughtfully evaluate the proposed retrieval method, four state-of-the-art retrieval frameworks proposed for histopathological images are compared. They are:

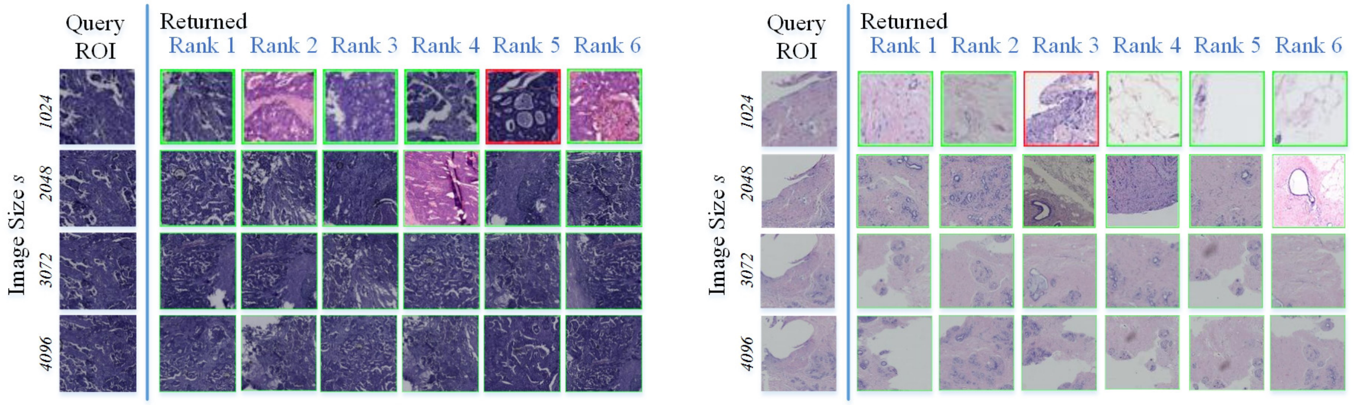
- *BoF-Cos* (Caicedo et al. [14]): The low-level features are quantified by BoF model to represent histopathological images and the similarity between images are measured by cosine distance.
- *LDA-Cos* (Ma et al. [24]): High-level semantic features generated by latent-Dirichlet-allocation (LDA-based) [31] are used to describe images and the similarity between images are measured by cosine distance.
- *SBC-KSH* (Zhang et al. [21]): The SIFT-BoF representations are converted to binary codes using KSH model [23] and the similarities are measured by hamming distance.
- *LDA-SH* (Ma et al. [32]): The low-level features are converted to binary codes by a LDA-based supervised hashing model and the similarities are measured by hamming distance.

In the four frameworks, BoF-Cos and LDA-Cos are established based on unsupervised algorithms. They are also designed for retrieval task that lacks labeled histopathological iamges. In contrast, SBC-KSH and LDA-SH are based on supervised hashing models, which are designed for retrieval situation that plenty of labeled images are available for training. To evaluate the performance of the proposed framework in the latter situation, we also implemented a MBC-KSH framework by replacing the hashing model tPCA with KSH. the results of MBC-KSH are also discussed in this section.

For fair comparison, the low-level features in the four compared frameworks are set the same SIFT descriptors. The parameters in each model are optimized in the training set and results are obtained in the testing set with optimal parameters. The MAP of the top 20 returned ROIs is used to evaluate the performance of retrieval. In addition, the lesion of the query ROI can be classified following the K-nearest neighbor paradigm when the ROIs in the retrieval database are labeled. In this experiment, the lesion of each testing sample is determined from the 20-nearest neighbors (the top 20 returned ROIs) and the mean classification accuracy of the testing samples is reported. Moreover, the classification performance of traditional classifiers, including linear SVM, KNN based on Euclidean distance and Softmax classifier, are provided as the classification benchmark.

TABLE III
AVERAGE PROPOSAL NUMBER, MAP AND AVERAGE TIME CONSUMPTION AS A FUNCTION OF BIT NUMBER K.

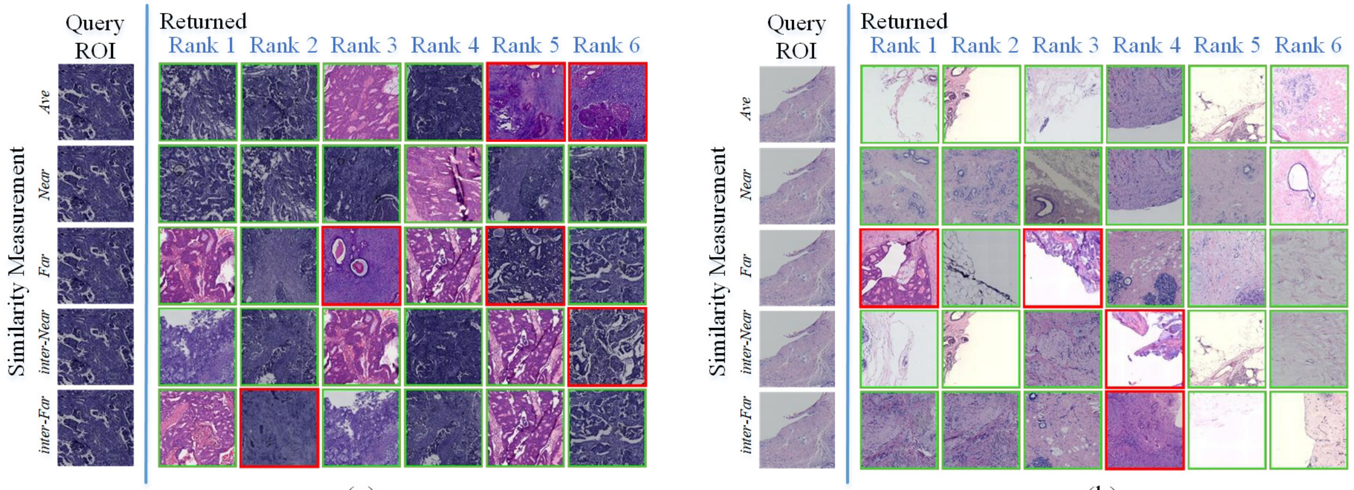
K	s=1024			s=2048			s=3072			s=4096		
	#Proposal	MAP	Time(s)	#Proposal	MAP	Time(s)	#Proposal	MAP	Time(s)	#Proposal	MAP	Time(s)
16	1254	0.823	0.022	4192	0.946	0.073	7704	0.952	0.324	11466	0.954	0.819
24	393	0.872	0.008	1346	0.944	0.029	2493	0.944	0.115	3874	0.952	0.289
32	100	0.914	0.017	560	0.942	0.028	872	0.941	0.042	973	0.948	0.101
40	64	0.941	0.003	381	0.954	0.012	667	0.956	0.047	952	0.959	0.101
48	**62**	**0.959**	**0.003**	**371**	**0.964**	**0.011**	**646**	**0.963**	**0.046**	**934**	**0.960**	**0.106**
56	62	0.960	0.003	371	0.965	0.012	646	0.963	0.051	930	0.960	0.138
64	62	0.960	0.003	371	0.966	0.011	646	0.963	0.040	930	0.960	0.125



(a)

(b)

Fig. 7. Instances of retrieval for scenes in different size, where (a) is the result for a scene of invasive carcinoma of no special type (NST), and (b) is the result for normal tissue. The first column of each instance shows the query ROIs and the other columns are results. The correct and incorrect retrieval results are framed in green and red, respectively.



(a)

(b)

Fig. 8. Retrieval for the two instances in Figure 7 by different similarity measurements for 2048 × 2048 query ROIs.

1) Results on Motic database: The division of training and testing set is the same as that in section III-D. Since the four comparative methods are all designed for database that consists of ROIs, we converted the retrieval database into four sub-databases of individual ROIs through randomly sampling the WSIs. The side length of ROIs in the four sub-databases is 1024, 2048, 3072, and 4096, with the number of ROIs 10290, 9734, 8784, and 8194, respectively. The results for the four sub-databases are given in Table IV.

Overall, the proposed method MBC-tPCA achieved the state-of-the-art retrieval performance in the four unsupervised methods for ROIs in 2048×2048, 3072×3072 and 4096×4096 pixels. And for ROIs in 1024×1024 pixels, LDA-Cos performed the best, the MAP of which is 1% higher to MBC-

TABLE IV
RETRIEVAL AND CLASSIFICATION PERFORMANCE OF THE COMPARED METHODS ON THE MOTIC DATABASE.

Unsupervised	s = 1024	s = 2048	s = 3072	s = 4096
Methods	MAP / MCA	MAP / MCA	MAP / MCA	MAP / MCA
BoF-Cos[14]	0.67 / 0.81	0.71 / 0.83	0.74 / 0.86	0.74 / 0.86
LDA-Cos[24]	**0.72** / **0.84**	0.78 / **0.88**	0.82 / 0.91	0.83 / 0.91
SBC-tPCA[27]	0.66 / 0.69	0.71 / 0.72	0.74 / 0.77	0.76 / 0.73
MBC-tPCA	0.71 / 0.76	**0.80** / 0.86	**0.86** / **0.91**	**0.88** / **0.92**

Supervised	s = 1024	s = 2048	s = 3072	s = 4096
Methods	MAP / MCA	MAP / MCA	MAP / MCA	MAP / MCA
SVM	- / 0.75	- / 0.77	- / 0.78	- / 0.80
KNN	- / 0.81	- / 0.83	- / 0.86	- / 0.86
Softmax	- / 0.81	- / 0.84	- / 0.88	- / 0.89
SBC-KSH[21]	0.75 / 0.76	0.85 / 0.85	0.88 / 0.88	0.90 / 0.89
LDA-SH[32]	**0.81** / **0.82**	**0.89** / 0.89	0.91 / 0.92	0.91 / 0.92
MBC-KSH	0.78 / 0.78	0.88 / **0.89**	**0.93** / **0.94**	**0.94** / **0.95**

TABLE V
RETRIEVAL AND CLASSIFICATION PERFORMANCE OF THE COMPARED METHODS ON THE BREAKHIS DATABASE.

Unsupervised	M. = 40×	M. = 100×	M. = 200×	M. = 400×
Methods	MAP / MCA	MAP / MCA	MAP / MCA	MAP / MCA
BoF-Cos[14]	0.37 / 0.45	0.33 / **0.42**	0.30 / 0.38	0.28 / 0.35
LDA-Cos[24]	0.39 / 0.43	**0.36** / 0.40	0.33 / 0.40	**0.31** / **0.38**
SBC-tPCA[27]	0.34 / 0.43	0.31 / 0.40	0.29 / 0.39	0.25 / 0.34
MBC-tPCA	**0.41** / **0.47**	0.35 / 0.40	**0.33** / **0.40**	0.27 / 0.37

Supervised	M. = 40×	M. = 100×	M. = 200×	M. = 400×
Methods	MAP / MCA	MAP / MCA	MAP / MCA	MAP / MCA
SVM	- / 0.39	- / 0.31	- / 0.28	- / 0.21
KNN	- / 0.42	- / 0.41	- / **0.37**	- / **0.36**
Softmax	- / 0.40	- / 0.31	- / 0.28	- / 0.21
SBC-KSH[21]	**0.41** / 0.41	0.33 / 0.30	0.30 / 0.25	0.27 / 0.22
LDA-SH[32]	0.40 / 0.39	**0.33** / 0.31	**0.31** / 0.25	0.27 / 0.24
MBC-KSH	0.38 / **0.43**	0.32 / **0.37**	0.29 / 0.36	**0.28** / 0.34

tPCA. The superiority of LDA-Cos derives from the high-level semantics generated by LDA. And, the advantage of MBC-based representations becomes larger when the size of ROI increases. Remarkably, the MAP of MBC-tPCA is 5% better than LDA-Cos when $s = 4096$. That is because larger ROIs contain more meaningful objects. Using local representations can better characters these objects than using a global representation, and thus obtains a higher accuracy in retrieval.

Utilizing the supervised hashing method, the proposed MBC-KSH achieved a better performance than the unsupervised model MBC-tPCA. Compared MBC-KSH with SBC-KSH, the retrieval accuracy improves more than 3%. These results have demonstrated that the proposed MBC-based similarity measurement is also effective for the supervised hashing model when the labels of images are accessible.

In addition, the classification performance of the proposed retrieval framework is also competitive in the compared methods. Especially for $s = 2048, 3072, 4096$, MBC-KSH achieves the best classification accuracy over all the compared methods. It indicates that the proposed framework is applicable to breast lesion classification.

2) Results on BreaKHis database: The comparison is also conducted on the public database BreaKHis [33], which contains eight classes of breast tumor images with four magnifications. Specifically, there are 1995, 2981, 2913, and 1820 images in magnification of 40×, 100×, 200× and 400×. The same experiment as on the Motic database is conducted on this database, where the division of training and testing data is the same with that in research [33]. The results of sub-lesion retrieval and classification for ROIs with the four magnifications are given in Table V.

The MBC-tPCA method performs 2%-7% better than SBC-tPCA in the retrieval task. Again, it proves the effectiveness of the multi-bianry-code-based similarity measurement. Notice that this superiority is more evident in lower magnification. This is because the images in BreakHis database are all in size of 700×460 pixels so that images in lower magnification contains more information for the lesion. In low magnifications, dividing the image into tiles to encode preserved more details of the lesion, and thereby achieved better retrieval performance. The conclusion is consistent with that reached in the Motic Database.

Utilizing the supervised model, MBC-KSH achieved rather worse performance than MBC-tPCA in the retrieval and classification task. One of the reason is that the tile-level labels used to train KSH are imprecise. Specifically, the tile-level labels inherited from the image-level labels, since only image-level labels are accessible in BreakHis database. In this situation, the tumor-irrelevant tiles in the image were also labeled as the tumor, which confused the KSH model in the training stage.

IV. CONCLUSION

In this paper, we proposed a novel content-based image retrieval framework for database that consists of histopathological whole slide images. The effectiveness of the framework has been certified with experiments on two breast tumor databases. The contribution of this work mainly includes the following two aspects. The one is that we have proposed a complete size-scalable CBIR framework for large-scale database that consists of WSIs. Using the binarization method and hashing technique, the query process can be efficiently completed. Another is that we have proposed a set of similarity measurement for the images that represented in multiple binary codes, in which the Nearest and inter-Nearest distances are certified effective for histopathological image retrieval. Further work will concentrate on the retrieval with irregular ROIs using the proposed similarity measurements.

REFERENCES

[1] S. R. Lakhani, I. A. for Research on Cancer, W. H. Organization *et al.*, *WHO classification of tumours of the breast.* International Agency for Research on Cancer, 2012.

[2] J. S. Duncan and N. Ayache, "Medical image analysis: Progress over two decades and the challenges ahead," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 22, no. 1, pp. 85–106, 2000.

[3] C. Mosquera-Lopez, S. Agaian, A. Velez-Hoyos, and I. Thompson, "Computer-aided prostate cancer diagnosis from digitized histopathology: A review on texture-based systems," *Biomedical Engineering, IEEE Reviews in*, vol. 8, pp. 98–113, 2014.

[4] M. N. Gurcan, L. E. Boucheron, A. Can, A. Madabhushi, N. M. Rajpoot, and B. Yener, "Histopathological image analysis: a review," *Biomedical Engineering, IEEE Reviews in*, vol. 2, pp. 147–171, 2009.

[5] D. Comaniciu, P. Meer, D. Foran, and A. Medl, "Bimodal system for interactive indexing and retrieval of pathology images," in *International Conference on Pattern Recognition*, 1998, pp. 76–81.

[6] D. Comaniciu, P. Meer, and D. Foran, "Shape-based image indexing and retrieval for diagnostic pathology," in *Pattern Recognition, 1998. Proceedings. Fourteenth International Conference on*, vol. 1. IEEE, 1998, pp. 902–904.

[7] A. W. Wetzel, R. Crowley, S. Kim, R. Dawson, L. Zheng, Y. M. Joo, Y. Yagi, J. Gilbertson, C. Gadd, and D. W. Deerfield, "Evaluation of prostate tumor grades by content-based image retrieval," *Proceedings of SPIE - The International Society for Optical Engineering*, vol. 3584, pp. 244–252, 1999.

[8] L. Zheng, A. W. Wetzel, J. Gilbertson, and M. J. Becich, "Design and analysis of a content-based pathology image retrieval system." *IEEE Transactions on Information Technology in Biomedicine*, vol. 7, no. 4, pp. 249–55, 2004.

[9] G. Zhou and L. Jiang, "Content-based cell pathology image retrieval by combining different features," *Medical Imaging Pacs and Imaging Informatics*, vol. 5371, pp. 326–333, 2004.

[10] N. Mehta, R. S. Alomari, and V. Chaudhary, "Content based subimage retrieval system for high resolution pathology images using salient interest points." in *International Conference of the IEEE Engineering in Medicine and Biology Society*, 2009, pp. 3719–22.

[11] S. Doyle, M. Hwang, N. S., F. MD., T. JE., and M. A., "Using manifold learning for content-based image retrieval of prostate histopathology." in *Medical Image Computing and Computer-Assisted Intervention*, 2007.

[12] R. Sparks and A. Madabhushi, "Out-of-sample extrapolation using semi-supervised manifold learning (ose-ssl): Content-based image retrieval for prostate histology grading," *International Symposium on Biomedical Imaging*, pp. 734–737, 2011.

[13] J. C. Caicedo, F. A. Gonzalez, and E. Romero, "A semantic content-based retrieval method for histopathology images," in *Asia Information Retrieval Conference on Information Retrieval Technology*, 2008, pp. 51–60.

[14] J. C. Caicedo and E. Izquierdo, "Combining low-level features for improved classification and retrieval of histology images," *Ibai Publishing*, vol. 2, no. 1, pp. 68–82, 2010.

[15] J. Shi, Y. Ma, Z. Jiang, H. Feng, J. Chen, and Y. Zhao, "Pathological image retrieval for breast cancer with plsa model," in *International Conference on Image and Graphics*, 2013, pp. 634–638.

[16] Y. Zheng, Z. Jiang, J. Shi, and Y. Ma, "Pathology image retrieval by block lbp based plsa model with low-rank and sparse matrix decomposition," in *Chinese Conference on Image and Graphics Technologies*. Springer, 2014, pp. 327–335.

[17] Y. Ma, S. Jun, Z. Jiang, and F. Hao, "Plsa-based pathological image retrieval for breast cancer with color deconvolution," *Proceedings of SPIE*, vol. 8920, no. 8, pp. 89 200L–89 200L–7, 2013.

[18] Y. Zheng, Z. Jiang, J. Shi, and Y. Ma, "Retrieval of pathology image for breast cancer using plsa model based on texture and pathological features," in *IEEE International Conference on Image Processing*, 2014, pp. 2304–2308.

[19] A. Sridhar, S. Doyle, and A. Madabhushi, "Boosted spectral embedding (bose): Applications to content-based image retrieval of histopathology," in *International Symposium on Biomedical Imaging*, 2011, pp. 1897–1900.

[20] S. Zhang and D. Metaxas, "Large-scale medical image analytics: Recent methodologies, applications and future directions," *Medical Image Analysis*, vol. 33, pp. 98–101, 2016.

[21] X. Zhang, W. Liu, M. Dundar, S. Badve, and S. Zhang, "Towards large-scale histopathological image analysis: Hashing-based image retrieval," *Medical Imaging, IEEE Transactions on*, vol. 34, no. 2, pp. 496–506, 2015.

[22] X. Zhang, H. Dou, T. Ju, J. Xu, and S. Zhang, "Fusing heterogeneous features from stacked sparse autoencoder for histopathological image analysis," *IEEE Journal of Biomedical and Health Informatics*, vol. 20, no. 5, pp. 1377–1383, 2015.

[23] W. Liu, J. Wang, R. Ji, and Y. G. Jiang, "Supervised hashing with kernels," in *Computer Vision and Pattern Recognition*, 2012, pp. 2074–2081.

[24] Y. Ma, Z. Jiang, H. Zhang, F. Xie, Y. Zheng, H. Shi, and Y. Zhao, "Breast histopathological image retrieval based on latent dirichlet allocation," *IEEE Journal of Biomedical and Health Informatics*, pp. 1–1, 2016.

[25] Y. Zheng, Z. Jiang, Y. Ma, H. Zhang, F. Xie, H. Shi, and Y. Zhao, "Content-based histopathological image retrieval for whole slide image database using binary codes," in *SPIE Medical Imaging*, 2017.

[26] B. Kulis and K. Grauman, "Kernelized locality-sensitive hashing," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 6, pp. 1092–1104, 2012.

[27] J. Wang, S. Kumar, and S. F. Chang, "Semi-supervised hashing for scalable image retrieval," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, Ca, Usa, 13-18 June*, 2010, pp. 3424–3431.

[28] Y. Gong, S. Lazebnik, A. Gordo, and F. Perronnin, "Iterative quantization: a procrustean approach to learning binary codes for large-scale image retrieval." in *Computer Vision and Pattern Recognition*, 2011, pp. 2916–2929.

[29] M. A. Carreira-Perpinan and R. Raziperchikolaei, "Hashing with binary autoencoders," in *Computer Vision and Pattern Recognition*, 2015, pp. 557–566.

[30] R. Kohavi *et al.*, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Ijcai*, vol. 14, no. 2, 1995, pp. 1137–1145.

[31] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[32] Y. Ma, Z. Jiang, H. Zhang, F. Xie, Y. Zheng, H. Shi, and Y. Zhao, "Proposing regions from histopathological whole slide image for retrieval using selective search," in *IEEE International Symposium of Biomedical imaging*, 2017, online.

[33] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, "A dataset for breast cancer histopathological image classification," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 7, pp. 1455–1462, 2016.