Content-based histopathological image retrieval for whole slide image database using binary codes

Zheng Yushan[a,b], Jiang Zhiguo[a,b], Ma Yibing[a,b], Zhang Haopeng[a,b], and Fengying Xie[a,b]

[a]Image Processing Center, School of Astronautics, Beihang University, Beijing, 100191, China
[b]Beijing Key Laboratory of Digital Media, Beijing, 100191, China

ABSTRACT

Content-based image retrieval (CBIR) has been widely researched for medical images. In application of histopathological images, there are two issues that need to be carefully considered. The one is that the digital slide is stored in a spatially continuous image with a size of more than 10k×10K pixels. The other is that the size of query image varies in a large range according to different diagnostic conditions. It is a challenging work to retrieve the eligible regions for the query image from the database that consists of whole slide images (WSIs). In this paper, we proposed a CBIR framework for the WSI database and size-scalable query images. Each WSI in the database is encoded and stored in a matrix of binary codes. When retrieving, the query image is first encoded into a set of binary codes and analysed to pre-choose a set of proposal regions from database using hashing method. Then a multi-binary-code-based similarity measurement based on hamming distance is designed to rank proposal regions. Finally, the top relevant regions and their locations in the WSIs along with the diagnostic information are returned to assist pathologists in diagnoses. The effectiveness of the proposed framework is evaluated in a fine-annotated WSIs database of epithelial breast tumours. The experimental results show that proposed framework is both effective and efficiency for content-based whole slide image retrieval.

Keywords: CBIR, histopathological image, WSI, binary code, Hash, breast cancer

1. BACKGROUND AND PURPOSE

The research of content-based image retrieval (CBIR) for histopathological images can date back to 1998.[1] Then, some researchers[2–5] utilized the classical image features to depict the histopathological images and achieved the retrieval for histopathological images in cellar level and sub-image level. In recent years, researchers applied the high-level feature extraction models, such as manifold learning,[6] semantic analysis,[7–9] spectral embedding, and etc., to the histopathological image representation, which has improved the performance of CBIR.

To satisfy the application for the database of massive histopathological images, Zhang et al.[10] introduced hashing method into the content-based histopathological image retrieval. Instead of using high dimensional feature vector, they encode each histopathological image into an array of binary code and store it within tens of bits. Then the similarities among histopathological images are measured by hamming distance, which can be calculated very fast using bit operations by computer. After ranking the similarities, the most similar images with the query image are returned as the retrieval result. However, this method will suffer from two issues when applied to the database that consists of massive whole slide images (WSI). First, the WSI is a spatially continuous image while the query image is usually a region of interest (ROI) that captured by the doctor from the WSI. To retrieve the similar regions in the same size with the ROI, every feasible sub-images of the WSIs in the database need to be considered. It is obviously not applicable to calculate the hamming distance between the query ROI with all the feasible regions in large scale database. Second, the the size of the query ROI varies greatly according to diagnostic requirement. When the query ROI has a large size, e.g. 4000×4000 pixels, representing the image using a single binary code will neglect the local information of the histopathological image.

This work is to tackle the two issues and design a complete CBIR framework for the database that consists of WSIs.

Further author information: (Send correspondence to Jiang Zhiguo and Zheng Yushan)
Jiang Zhiguo : E-mail: jiangzg@buaa.edu.cn, Telephone: +86 010 82316173
Zheng Yushan: E-mail: yszheng@buaa.edu.cn, Telephone: +86 010 82338061

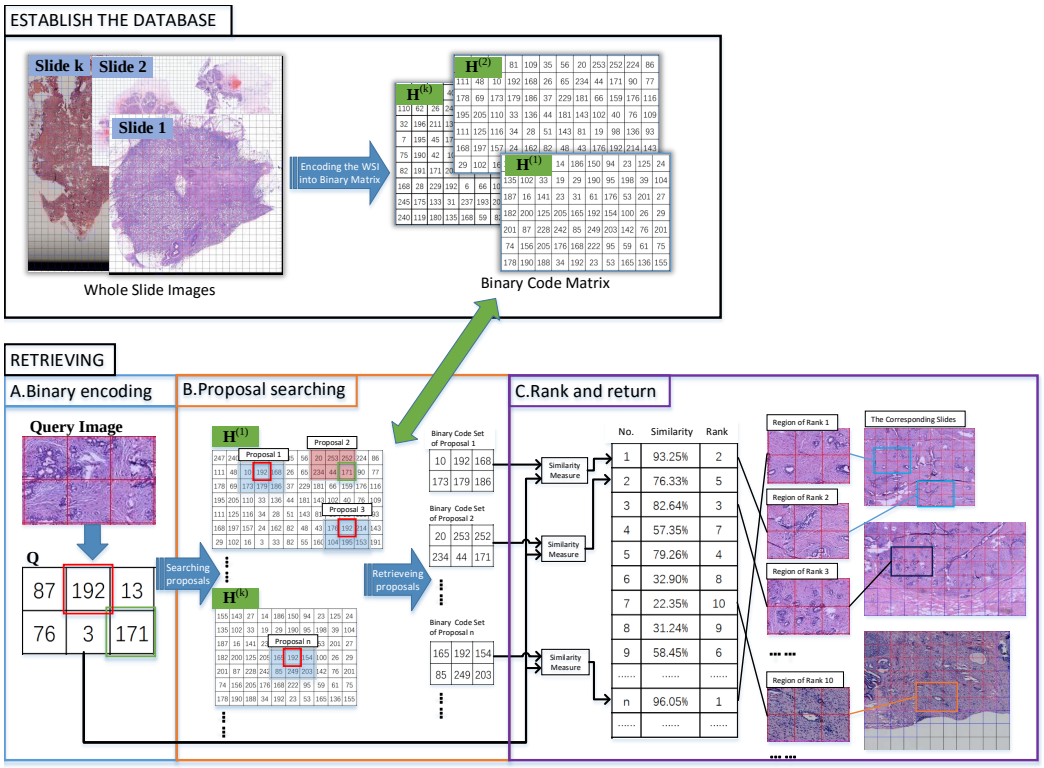


Figure 1. The flowchart of the proposed retrieval framework. For clearness, the binary code is displayed in decimal numeral.

2. METHOD

In this paper, we proposed a novel content-based retrieval framework for the whole slide images. The flowchart of the proposed framework is shown in Figure 1. For establishing the retrieval database, we divided the whole slide image into square tiles and represented each tile by a binary code, instead of sampling sub-images from the WSI. The retrieval procedure can be divided into 3 steps:

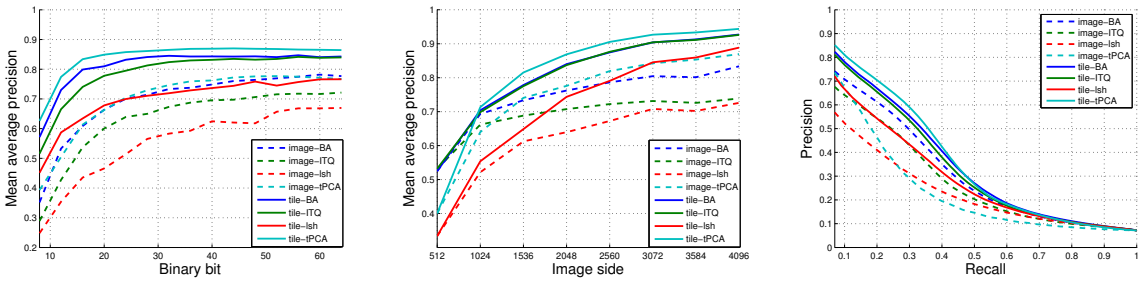
A. **Binary Encoding:** The query ROI is first divided into square tiles that have the same size with those in the database. Then, these tiles are encoded into a set of binary codes.

B. **Proposal Searching:** A group of proposal regions are first retrieved by searching the tiles that have the equivalent binary code with the tile in the query ROI. For clearness, the binary code is displayed in decimal numeral in Figure 1. This search can be completed very fast with a hash table that is established beforehand.

C. **Rank and Return:** The proposal regions are ranked using a similarity measurement based on multiple hamming distance. Specifically, let $\mathbf{Q} = \{q_1, q_2, ..., q_l\}$ denote the set of binary codes of the query ROI, $\mathbf{P} = \{p_1, p_2, ..., p_l\}$ denote the set of binary codes of a proposal region and l be the number of tiles related to the query image size. the the similarity measurement between Q and P is defined as

$$D(Q, P) = \sum_{i=1}^{l} \min_{j=1,2,...,l} hamming(q_i, p_j), \tag{1}$$

where, $hamming(q_i, p_j)$ is the hamming distance between the binary code q_i and p_j. As definition, the more similar the Q and P are, the smaller the $D(Q, P)$ will be. Finally, the top-ranked regions with their locations in the corresponding WSIs are returned as the retrieval result.



(a) Retrieval performance for various number of binary bits.

(b) Retrieval performance for various sizes of the query image.

(c) Precision-Recall curve at binary bit = 28 & image side = 2048.

Figure 2. Experimental results for individual images.

3. EXPERIMENT

3.1 Experimental Data

The experimental images used in this research are supplied by Motic (Xiamen) Medical Diagnostic Systems Co. Ltd.* There are 145 HE-stained whole slide images (WSIs) of 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, in which 50 WSIs are Completely annotated and the others are locally annotated. These annotations are labeled as 15 sub-categories of epithelial breast tumors according to the world health organization (WHO) standard.[11] Note that, in the query experiment, only the images that have the same label with the query image are regarded as the correct retrieval results. The features used in our research is bag of features (BoF) based on SIFT descriptors, which is certified effective in content-based histopathological image retrieval.[8,10] All the methods are implemented using MATLAB 2013a on a PC with a CPU of 2.10 GHz.

3.2 Results

First, we conducted experiment to certify the effectiveness of the similarity measurement defined by Equation 1 using individual images that sampled from the annotated regions. The mean average precision of the top 20 returned images is used as the assessment criteria. To evaluate the universality of our framework, 4 unsupervised binarization methods are validated. They are Kernelized Locality-Sensitive Hashing (LSH),[12] thresholded PCA (tPCA), Iterative Quantization (ITQ)[13] and Binary Autoencoders (BA).[14] The results for different size of query image and different number of bits of the binary code are given in Figure 2, in which the 'tile-' prefixed curves are the retrieval results using the proposed similarity measurement. The retrieval results based on the image-level binary code (used in[10]) are prefixed by 'image-'. The results show that the proposed similarity measurement is effective and superior to the present method. In the 4 binarization methods, tPCA achieved the best performance.

Choosing tPCA as the binarization method, we conducted experiments to evaluate the effectiveness of the proposed framework for WSI database. In this experiment, the size of the query image is evaluated from 512×512 to 4096×4096. The 50 completely annotated WSIs are used to establish the database. And for each testing size, 6600 images sampled from the annotated regions in the other WSIs are used as the query images. The mean average precisions (mAP) and average retrieval times (ART) for different numbers of returned images R are shown in Table 1. It shows that the proposed framework is both effective and efficient.

Table 1. Retrieval performance in the database of WSIs

	R = 10				R = 20				R = 30			
Image Side	1024	2048	3072	4096	1024	2048	3072	4096	1024	2048	3072	4096
mAP	0.886	0.911	0.941	0.944	0.879	0.839	0.918	0.930	0.877	0.801	0.890	0.917
ART (ms)	0.96	3.5	5.9	10.4	0.96	3.5	5.9	10.4	0.96	3.5	5.9	10.4

*Motic (Xiamen) Medical Diagnostic Systems Co. Ltd., Xiamen 361101, China

4. CONTRIBUTIONS AND CONCLUSION

There are two contributions of our work. The one is that we have proposed a complete size-scalable CBIR framework for large scale database of WSIs. Using the binarization method and hashing technique, the query process can be completed vary fast. Specifically for the large size query image of 4096×4096, the retrieval precision of the top 20 return is 93% and average retrieval time is about 10 millisecond. The second is that we proposed a similarity measurement for the images that represented in multiple binary codes. The main operand of proposed method is from the ranking step, which varies with the number of proposal regions. Choosing the appropriate number of binary bits according to the amount of WSIs in the dataset can effectively control proposal number and, hence, reduce the operand in retrieval, contributing to a relatively constant querying time for large scale WSI database.

5. ACKNOWLEDGMENT AND DECLARATION

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

REFERENCES

[1] Comaniciu, D., Meer, P., Foran, D., and Medl, A., "Bimodal system for interactive indexing and retrieval of pathology images," in [*Applications of Computer Vision, 1998. WACV '98. Proceedings., Fourth IEEE Workshop on*], 76–81 (1998).

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

[3] Zheng, L., Wetzel, A. W., Gilbertson, J., and Becich, M. J., "Design and analysis of a content-based pathology image retrieval system.," *IEEE Transactions on Information Technology in Biomedicine* **7**(4), 249–55 (2004).

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

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

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

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

[8] Caicedo, J. C. and Izquierdo, E., "Combining low-level features for improved classification and retrieval of histology images.," *Ibai Publishing* **2**(1), 68–82 (2010).

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

[10] Zhang, X., Liu, W., Dundar, M., Badve, S., and Zhang, S., "Towards large-scale histopathological image analysis: hashing-based image retrieval.," *IEEE Transactions on Medical Imaging* **34**(2), 496–506 (2015).

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

[12] Kulis, B. and Grauman, K., "Kernelized locality-sensitive hashing," *Pattern Analysis and Machine Intelligence IEEE Transactions on* **34**(6), 1092–1104 (2012).

[13] Gong, Y., Lazebnik, S., Gordo, A., and Perronnin, F., "Iterative quantization: a procrustean approach to learning binary codes for large-scale image retrieval.," in [*Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*], 2916–2929 (2011).

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