RETRIEVAL OF PATHOLOGY IMAGE FOR BREAST CANCER USING PLSA MODEL BASED ON TEXTURE AND PATHOLOGICAL FEATURES

Yushan Zheng Zhiguo Jiang Jun Shi Yibing Ma

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

ABSTRACT

Content-based image retrieval (CBIR) for digital pathology slides is of clinical use for breast cancer aided diagnosis. One of the largest challenges in CBIR is feature extraction. In this paper, we propose a novel pathology image retrieval method for breast cancer, which aims to characterize the pathology image content through texture and pathological features and further discover the latent high-level semantics. Specifically, the proposed method utilizes block Gabor features to describe the texture structure, and simultaneously designs nucleus-based pathological features to describe morphological characteristics of nuclei. Based on these two kinds of local feature descriptors, two codebooks are built to learn the probabilistic latent semantic analysis (pLSA) models. Consequently, each image is represented by the topics of pLSA models which can reveal the semantic concepts. Experimental results on the digital pathology image database for breast cancer demonstrate the feasibility and effectiveness of our method.

Index Terms—Image retrieval, feature extraction, computer aided diagnosis, breast cancer, probabilistic latent semantic analysis

1. INTRODUCTION

Digital pathology slide has been widely concerned in the last decades. A great many of companies and universities, such as Leica, Motic, Definiens, University of Leeds and University of Pittsburgh Medical Center (UPMC), have focused on pathology image analysis and also built pathology slide databases for aiding pathologists during the diagnosis process through retrieving similar previously diagnosed cases.

Based on these slide databases, many Computer Aided Diagnosis (CAD) systems for different types of cancer are established to improve the accuracy of diagnosis. Specifically, CAD for breast cancer has attracted more attention due to its high incidence in female cancer cases [1, 2]. In the past years, new technologies for breast cancer diagnosis have developed rapidly. Yet the final diagnosis of breast cancer still depends on the pathological methods [3] and the most important factor that affects the level of pathologist is clinical experience. CAD system consisting of pathology slide database with confirmed diagnosis information can well support pathologists. However, the database usually contains massive amounts of slides with much higher resolution than common digital image. Therefore, CAD systems that can effectively retrieve useful cases from big pathology image data to support the diagnosis process are urgently required.

To enhance the retrieval performance of CAD, Content-Based Image Retrieval (CBIR) has been proposed and successfully applied to many clinical applications [4, 5]. Particularly, feature extraction is of critical importance for CBIR, which can accurately describe the image content by a meaningful low dimensional representation. Over the past years, many pathological feature extraction methods for CBIR have been developed. Caicedo et al. [6] apply different kinds of visual features to achieve the retrieval task for four kinds of tissues. Recently, Kowal et al. [7] have paid more attention to statistical features of individual nuclei to classify benign and malignant cases of breast cancer. Obviously these methods mentioned above just describe the image content from one way (visual features or statistical features of nuclei) and may even ignore the high-level semantic concepts that may exist in pathology image.

In this paper, we present a novel retrieval method of pathology images for breast cancer, which takes both local Gabor features and nucleus-based pathological features as the low-level features and then applies probabilistic latent semantic analysis (pLSA) [8] model to discover the high-level semantics. Following our previous work [9], the entire pathology image is divided into non-overlapping blocks and

This work was supported by the National Natural Science Foundation of China (No. 61371134) and the 973 Program of China (Project No. 2010CB327900).

then Gabor features of each block under different scales and orientations are used to describe spatial texture variations which are likely to reflect some characteristics of breast cancer (e.g., various types of cellular atypia, different aspects of cell polarity and varying extents of infiltrative growth). Note that Scale Invariant Feature Transform (SIFT) descriptors after saliency detection are also used as low-level features in our prior work [9]. However, these features only characterize the image content in terms of visual attention and thus fail to reveal the pathological features. Therefore, in this paper, we also develop nucleus-based pathological features. Concretely, Retinex processing [10] is used for image enhancement and color normalization. Then color deconvolution [11] and Otsu method [12] are applied to extract nuclei. Afterwards, the statistical features of each nucleus (e.g., nuclear size, shape, and regularities of distribution) will be computed, which are denoted as the nucleus-based pathological features and further can reflect morphological characteristics of the nucleus. Based on these two kinds of local feature descriptors, two codebooks are built through *k*-means clustering and thus two pLSA models can be learnt. Finally each image can be represented by the combination of topics from these two pLSA models. Experimental results on digital pathology image database containing five kinds of breast cancer demonstrate the feasibility and effectiveness of our method.

The rest of this paper is arranged as follows: Section 2 introduces low-level feature description of our method. Section 3 describes high-level semantic representation using pLSA. Section 4 presents the pathology image database and experimental results. Finally Section 5 gives the conclusion.

2. LOW-LEVEL FEATURE DESCRIPTION

2.1. Local Gabor texture feature

As Gabor features [13] can detect texture variations under different scales and orientations, we use the Gabor filter responses with 4 scales and 8 orientations to describe texture information of pathology image. To further discover the spatial locality of texture structure, we divide the entire image (256×256) into non-overlapping blocks, and then extract Gabor features of each block (32×32). Consequently, there will be 32 Gabor response images for each block, and the mean and standard deviation of each response image are regarded as the features under specific scale and orientation. Finally we can obtain a 64 dimensional feature vector to characterize the texture information of each block.

2.2 Nucleus-based pathological feature

According to pathology, the nuclei and cytoplasm are generally stained by different colors. For example, in Fig. 1(a), the pathology images of breast cancer are dyed with hematoxylin and eosin (HE). However, as can be seen in

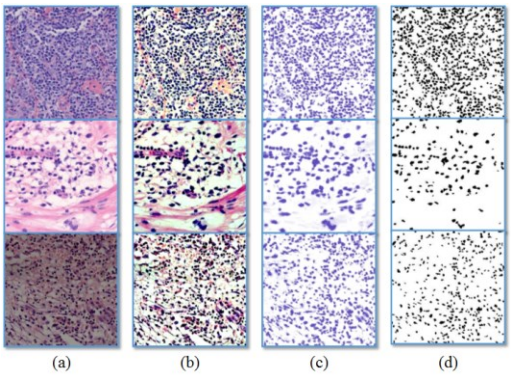

Fig. 1 (a) Pathology images of breast cancer stained by HE. (b) Retinex processing. (c) Nuclei separated by color deconvolution. (d) Segmentation results by global optimal threshold.

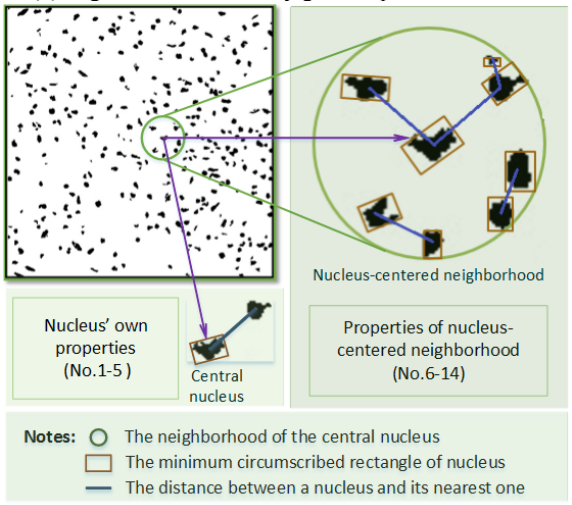

Fig. 2 The 14 dimensional features of a nucleus.

Table 1 Meaning of 14 dimensional features

No	Nucleus' own properties
1	Area (number of pixels)
2	Mean of gray-level after Retinex processing
3	Standard deviation of gray-level after Retinex processing
4	The length-width ratio of minimum circumscribed rectangle (width / length)
5	The distance between the nucleus and its nearest one
	Properties of nucleus-centered neighborhood
6	Number of the nuclei
7	Mean of nuclei areas
8	Standard deviation of nuclei areas
9	Mean of length-width ratios of minimum circumscribed rectangles for nuclei
10	Standard deviation of length-width ratios of minimum circumscribed rectangles for nuclei
11	Mean of distances between the central nucleus and other nuclei
12	Standard deviation of distances between the central nucleus and other nuclei
13	Mean of distances between each nucleus and its nearest one
14	Standard deviation of distances between each nucleus and its nearest one

Fig.1(a), the slides usually vary significantly due to the staining skill, smear preparation and the imaging condition. Consequently, the brightness and contrast between the slides are greatly different, which may influence the segmentation effect when extracting nuclei for quantitative analysis.

To deal with this problem, we firstly apply Retinex processing [10] for image enhancement and color normalization. As a result, the hues of the pathology images turn to be consistent and simultaneously nuclei seem to stand out from the cytoplasm, as shown in Fig. 1(b). Then we utilize color deconvolution [11] to separate different stain components and thus obtain the nuclei regions in Fig. 1(c). Considering the nuclei have the consistent color mode and stand out against the background after color deconvolution, the Otsu method [12] is used to segment the nuclei accurately. The results can be seen in Fig. 1(d).

To quantitatively analyze nuclei, connected component analysis is performed on the pathology images segmented by Otsu method. Consequently, small connected regions are removed. For the remaining regions, we design a 14 dimensional feature vector to characterize the properties of each nucleus and its neighborhood, whose radius is set as twice the distance between central nucleus and its nearest one. It is exhibited in Fig. 2 and Table 1.

3. HIGH-LEVEL SEMANTIC REPRESENTATION

Although two kinds of local features mentioned above can effectively characterize the image content, they fail to precisely describe high-level semantic concepts existed in the pathology image. Bag-of-Features (BoF) [14] can narrow down the gap between the low-level features and high-level semantics. However, it is usually affected by the synonyms of visual words and thus fails to reveal the semantics among words. pLSA model [8] can reveal topical similarities among words and meanwhile avoid the polysemy of words. More importantly, it has lower computational cost than other topic model (e.g., Latent Dirichlet Allocation (LDA) [15]).

Given a collection of documents $\mathbf{D} = \{d_1, d_2, \ldots, d_M\}$ with a set of words $\mathbf{W} = \{w_1, w_2, \ldots, w_N\}$. Commonly, low-level features can be modeled as words and images are regarded as documents. Let $\mathbf{Z} = \{z_1, z_2, \ldots, z_T\}$ be the set of latent topics, which are viewed as the latent variables between words and documents. pLSA can be given as a maximum log-likelihood formulation [8] :

$$\begin{aligned} L &= \sum_{i=1}^{N}\sum_{j=1}^{M} n(d_i, w_j)\log P(d_i, w_j) \\ &= \sum_{i=1}^{N} n(d_i)\left[\log P(d_i) + \sum_{j=1}^{M}\frac{n(d_i, w_j)}{n(d_i)}\log\sum_{k=1}^{T} P(w_j \mid z_k)P(z_k \mid d_i)\right], \end{aligned} \quad (1)$$

where $P(d_i, w_j) = P(d_i)P(w_j \mid d_i)$, $P(w_j \mid d_i) = \sum_{k=1}^{T} P(z_k \mid d_i)P(w_j \mid z_k)$, $n(d_i, w_j)$ represents the frequency that word w_j occurs in document d_i and $n(d_i)$ denotes the occurrence frequency of d_i. The goal of pLSA is to seek the optimal $P(z_k|d_i)$ and $P(w_j|z_k)$ through expectation-maximization (EM) algorithm [8], and the $P(z|d_i)$ is the topic representation of the i-th document.

The workflow chart of our method is given in Fig. 3. In the training stage, both nucleus-based pathological features and local Gabor features are extracted. Then two codebooks can be gained through k-means and thus the word-level representation corresponding to each image can be obtained through vector quantization, namely $P(w|d)$. EM algorithm is used to compute the optimal $P(z|d)$ and $P(w|z)$ in Eq. (1) and $P(z|d)$ is the topic representation of each image. Finally the two topic representations are combined as the final representation. In the test stage, the input ROI will be represented by the topics of two trained pLSA models. After computing the similarities between ROI and the images stored in the database, the top R similar images along with the confirmed diagnosis information are returned.

4. EXPERIMENT

The experiment is conducted on the pathology image database for breast cancer with confirmed diagnosis information, which is from Motic digital slide database for the yellow race[1]. The image database contains 5 categories and 600 images (256×256, 20x magnification) for each category, as shown in Fig. 4. Note that 50 images of each category are used for training and the remaining for test.

For each test sample, the top R=20 similar images are returned to evaluate the retrieval precision:

$$precision = \sum_{i=1}^{c} n_i / (C \times R), \quad (2)$$

where n_i is the number of returned images that have the same label with ROI and C is the number of test samples.

Considering different numbers of words and topics will affect the performance of pLSA, we select the optimal word number N from 20 to 200 and topic number T from 5 to 15. Specifically, N is set to 150 and T is set to 12. Furthermore, 4 distance measurements are used to compute the similarities between ROI and the images stored in the database. Table 2 demonstrates the retrieval precision of different methods. Note that Nucleus-based pLSA employs the pathological features proposed by ours for pLSA training, Gabor-based pLSA applies the local Gabor features proposed by ours, and Nucleus-Gabor-based BoF means that both nucleus and Gabor features are used for BoF. From Table 2, we can clearly see that our method outperforms other methods under different similarity measurements. Particularly the precision under cosine distance is optimal and up to 94.4%. Compared with the method proposed by Kowal et al. [7], Nucleus-based pLSA has superior retrieval performance. It is likely because the nucleus-based pathological features designed by ours have a better ability to characterize the local morphological properties of pathology image and

[1] http://www.mpathology.cn/Category_112/Index.aspx

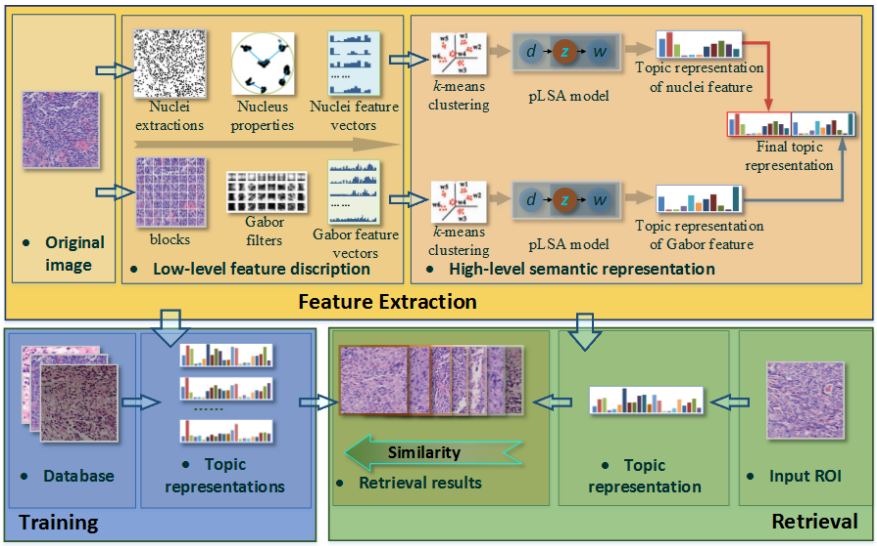

Fig. 3 The workflow chart of our retrieval framework.

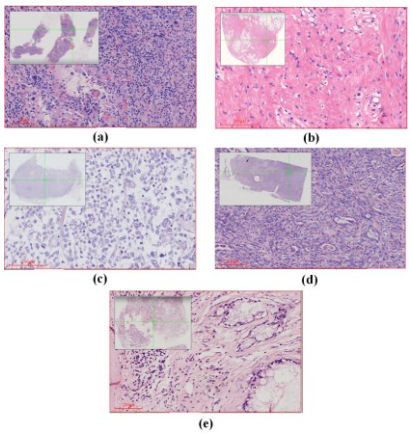

Fig. 4 5 categories of digital pathology slides. (a) Basal-like carcinoma (BLC). (b) Breast myofibroblastoma (BMFB). (c) Invasive breast cancer (IBC). (d) Low-grade adenosquamous carcinoma (LGASC). (e) Mucinous cystadenocarcinoma (MCA).

Table 2 Precisions (%) at the top 20 returns of seven methods

Methods	Euclidean distance	Cosine distance	Chi-square distance	Histogram intersection
Kowal et al. [7]	58.2	59.3	58.9	60.3
Caicedo et al. [6]	62.5	63.3	63.1	60.9
Nucleus-based pLSA	83.5	86.9	85.3	82.9
Gabor-based pLSA	87.4	87.6	87.2	87.6
Nucleus-Gabor-based BoF	89.8	90.4	90.3	89.5
SIFT-Gabor-pLSA [9]	88.4	88.4	86.8	85.9
Our method	**92.1**	**94.4**	**93.1**	**91.9**

simultaneously it may benefit from the high-level semantics of pLSA. The method proposed by Caicedo et al. [6] utilizes different kinds of visual features,however, Gabor-based pLSA is superior due to its local feature description and high-level semantics. Note that Nucleus-Gabor-based BoF is better than Nucleus-based or Gabor-based pLSA due to the contribution of feature combination. In contrast with Nucleus-Gabor-based BoF, our method is more effective since it overcomes the limits of BoF. And more remarkable, our method has greatly improved compared with our prior work (i.e., SIFT-Gabor-pLSA [9]), because it focus much on pathological features.

Fig. 5 shows the confusion matrix of our method under cosine distance. The diagonal elements represent the retrieval precisions and the others are confused retrieval ratios. Obviously our method has excellent retrieval performance for these 5 categories. Note that the confusion degree between BLC and LGASC is high since they are similar in terms of nuclear morphology and texture structure.

5. CONCLUSION

In this paper, we propose a retrieval method of pathology image for breast cancer using pLSA model based on texture and pathological features. It uses local Gabor the features to characterize texture structure, and simultaneously designs nucleus-based pathological features to characterize morphological properties of nuclei. Then it applies these two feature descriptors to train two pLSA models, which can describe the semantic concepts. Finally each image is represented by the combination of topics from these two pLSA models. Experimental results demonstrate the effectiveness of our method. Further research will aim to apply deep learning to automatically learn the feature representation of pathology image and simultaneously use parallel computation to improve the efficiency of training.

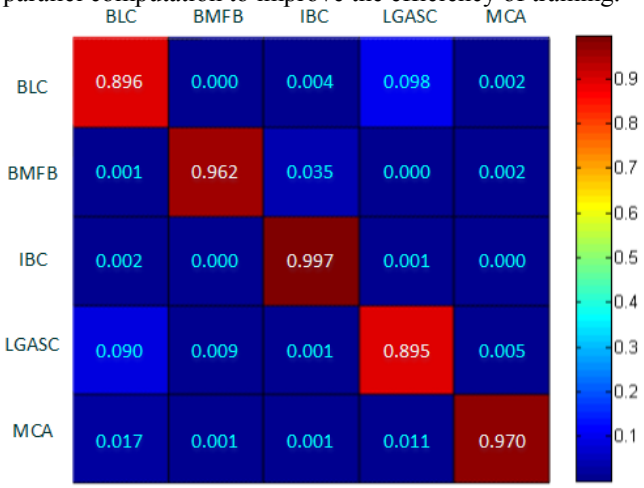

Fig. 5 The confusion matrix of our method.

6. REFERENCES

[1] Rebecca Siegel, Deepa Naishadham, and Ahmedin Jemal, "Cancer Statistics, 2013," *CA Cancer journal for Clinicians*, 63(1), pp. 11-30, January 2013.

[2] N. Li, R.S. Zheng, S.W. Zhang, X.N. Zou, H.M. Zeng, Z. Dai, and W.Q. Chen, "Analysis and Prediction of Breast Cancer Incidence Trend in China," *Chinese Journal of Preventive Medicine*, 46(8), pp. 703-707, 2012.

[3] Xilin Fu, *The Atlas for Pathologic Diagnosis of Breast Tumours,* Scientifics and Technical Documents Publishing House, Beijing, China, July 2013.

[4] Zhiyun Xue, L. Rodney Long, Sameer Antani, and George R. Thoma, "Pathological-based Vertebral Image Retrival," *IEEE International Symposium on Biomedical Imaging*: *From Nano to Macro (ISBI)*, Chicago, pp. 1893-1896, March 30-April 2, 2011.

[5] Lijia Zhi, Shaomin Zhang, Dazhe Zhao, Hong Zhao, and Shukuan Lin, "Medical Image Retrieval Using Sift Feature," *IEEE 2nd International Congress on Image and Signal Processing*, Tianjin, pp. 1-4, October 17-19, 2009.

[6] J.C. Caicedo and E. Izquierdo, "Combining Low-level Features for Improved Classification and Retrieval of Histology Images," *Transactions on Mass-Data Analysis of Images and Signals*, 2(1), pp. 68-82, September 2010.

[7] Marek Kowal, Paweł Filipczuk, Andrzej Obuchowicz, Józef Korbicz, and Roman Monczak, "Computer-aided Diagnosis of Breast Cancer Based on Fine Beedle Biopsy Microscopic Images," *Computers in Biology and Medicine*, 43(10), pp. 1563-1572, August 2013.

[8] T. Hofmann, "Probabilistic Latent Semantic Analysis," *The 22nd Annual ACM Conference on Research and Development in Information Retrieval*, San Francisco, pp. 289-296, 1999.

[9] J. Shi, Y. Ma, Z.G. Jiang, H. Feng, J. Chen and Y. Zhao, "Pathological Image Retrieval for Breast Cancer with pLSA Model," *IEEE International Conference on Image and Graphics (ICIG)*, Qingdao, pp. 634-638, July 2013.

[10] B. Funt, F. Ciurea, and J. McCann, "Retinex in matlab," *Journal of the Electronic Imaging,* 13(1), pp. 48–57, Jan. 2004.

[11] A.C. Ruifrok and D.A. Johnston, "Quantification of Histochemical Staining by Color Deconvolution," *Analytical and Quantitative Cytology and Histology*, 23(4), pp. 291-299, 2001.

[12] N. Otsu, "A Threshold Selection Method from Gray-level Histograms," *IEEE Transactions on Systems, Man and Cybernetics*, 9(1), pp. 62-66, January 1979.

[13] B.S. Manjunath and W.Y. Ma, "Texture Features for Browsing and Retrieval of Image Data," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 18(8), pp. 837-842, 1996.

[14] L. Fei-Fei and P. Perona, "A Bayesian Hierarchical Model for Learning Natural Scene Categories," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, San Diego, pp. 524-531, June 20-25, 2005.

[15] D.M. Blei, A.Y. Ng, and M.I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research,* 3(4-5), pp. 993-1022, 2003.