Pathology Image Retrieval by Block LBP Based pLSA Model with Low-Rank and Sparse Matrix Decomposition

Yushan Zheng, Zhiguo Jiang, Jun Shi, and Yibing Ma

Image Processing Center, School of Astronautics, Beihang University
Beijing Key Laboratory of Digital Media
Beijing, China
yszheng@sa.buaa.edu.cn jiangzg@buaa.edu.cn
chris.shi331@gmail.com hemp110@126.com

Abstract. Content-based image retrieval (CBIR) is widely used in Computer Aided Diagnosis (CAD) systems which can aid pathologist to make reasonable decision by querying the slides with diagnostic information from the digital pathology slide database. In this paper, we propose a novel pathology image retrieval method for breast cancer. It firstly applies block Local Binary Pattern (LBP) features to describe the spatial texture property of pathology image, and then use them to construct the probabilistic latent semantic analysis (pLSA) model which generally takes advantage of visual words to mine the topic-level representation of image and thus reveals the high-level semantics. Different from conventional pLSA model, we employ low-rank and sparse matrix composition for describing the correlated and specific characteristics of visual words. Therefore, the more discriminative topic-level representation corresponding to each pathology image can be obtained. Experimental results on the digital pathology image database for breast cancer demonstrate the feasibility and effectiveness of our method.

Keywords: Image retrieval, computer aided diagnosis, breast cancer, probabilistic latent semantic analysis, low-rank and sparse matrix composition.

1 Introduction

Computer Aided Diagnosis (CAD) system for breast cancer has attracted more and more attention due to morbidity increase of breast cancer in female [1, 2]. Although new technologies for breast cancer diagnosis have developed rapidly in the past few years, the final diagnosis still relies on the pathological theories [3]. As the digital pathology slides spread, senior pathologists can mark the lesion area with detailed descriptions on the digital slides and share it to others through CAD systems or the Internet. In the other hand, junior pathological doctors can get valuable suggestions by searching slides that contain diagnosis information when facing indeterminable cases. Therefore, CAD systems consisting of pathology slide database with confirmed diagnosis information are urgently required. But it is challenging to retrieve useful slides from the enormous database effectively and accurately for the reason that the resolution of digital pathology

T. Tan et al. (Eds.): IGTA 2014, CCIS 437, pp. 327–335, 2014.

slide is usually much higher than common digital image and the characteristics of pathology image are much different from natural images.

To deal with the retrieval problem on digital pathology slide databases, Content-Based Image Retrieval (CBIR) has been proposed and successfully applied to clinical diagnosis [4, 5]. Over the past years, a large number of retrieval methods for pathology image have been proposed. Caicedo et al. [6] apply different kinds of visual features to achieve the retrieval task for four kinds of tissues. Kowal et al. [7] take advantage of statistical features of individual nuclei to classify benign and malignant cases of breast cancer. However, these methods mentioned above just describe the global characteristics of the image and may even ignore the high-level semantics that exist in the image. Therefore, to mine the texture information and local property of pathology image, we propose to divide the entire image into non-overlapping blocks and extract Local Binary Pattern (LBP) [8] descriptor in each block. Then LBP descriptors are used to build the codebook composed of visual words through k-means. Afterwards, each pathology image can be represented by the word frequency histogram via Bag-of-Words (BoW) [12] scheme. However, there are synonyms among visual words and thus make the word-level representation hard to discriminatively reveal the semantics in images. Therefore, probabilistic latent semantic analysis (pLSA) [9] model is applied in our method to mine the high-level semantics of words. Nonetheless, pLSA model just uses BoW scheme to discover the word distribution, which is likely to ignore that there are some correlated and specific characteristics among words. Consequently, the word-level representation of conventional pLSA model may fail to describe the image content precisely. To improve the discriminant ability of pLSA, we apply low-rank and sparse matrix decomposition technique [10, 13] to decompose the word-level representation into two meaningful parts (i.e., correlated and specific word-level representations), and then utilize them to train two pLSA models. Finally each image can be represented by the combination topics learned from these two models.

Our proposed method consists of two contributions. First, we use block LBP features to describe the spatial texture information and then apply pLSA model to discover the high-level semantics of pathology images. The second is that we use the low-rank and sparse matrix decomposition to obtain two word-level representations which can characterize the correlated and specific parts of the visual word distribution. As a consequence, the discriminant ability of word-level representation has been greatly improved. Experimental results on the digital pathology image database of breast cancer demonstrate the feasibility and effectiveness of our method.

The rest of the paper is arranged as follows: Section 2 introduces block LBP descriptor. Section 3 describes the pLSA model along with the low-rank and sparse matrix decomposition. Section 4 presents the experimental database and results. Finally the conclusion is given in Section 5.

2 Block Local Binary Pattern (LBP)

Local binary Pattern (LBP) [8] is a powerful local texture descriptor with the advantages of rotation invariance and orientation invariance. The pattern of each pixel is calculated by quantifying pixels of its neighborhood into a string of binary code. Generally, the size of neighborhood is defined to 3×3. The basic LBP code of the central pixel is computed as

$$LBP(p_c) = \sum_{i=0}^{7} 2^i b(g(p_i) - g(p_c)) \tag{1}$$

,

where p_c is the central pixel and p_i is the neighbor pixel of p_c, $g(p)$ is the gray value of p and $b(u)$ is the binary function that $b(u)$=1 if u is greater or equal to 0; $b(u)$=0 otherwise. The process of LBP is given in Fig. 1.

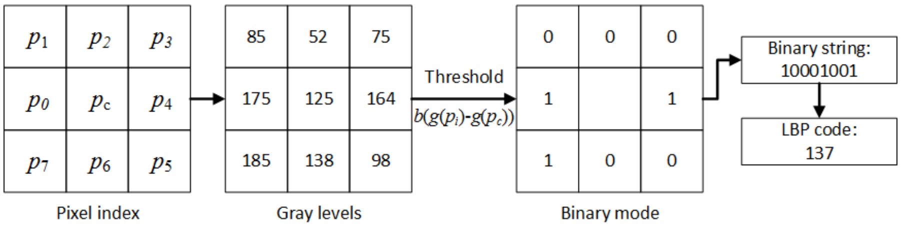

Fig. 1. The process of LBP

It can be found that the image is represented by a 256-dimensional (8-bits binary codes stand for 256 numbers in total) histogram by counting the pixel number of each LBP code. However, the 256-dimensional representation is redundancy and only 58 codes reflect the primitive structural information such as edges and corners. Then the dimension of the histogram is usually reduced to 59 by assigning non-uniform patterns to single bin [11].

It is obvious that LBP histogram or its uniform pattern is a global texture descriptor. To further discover the local structures, we divide the entire image into the non-overlapping blocks (16×16) and then compute uniform pattern of LBP in each block. Finally, pathology images can be represented by a sequence of 59-dimensional LBP histograms.

3 High-Level Semantic Mining

3.1 Probabilistic Latent Semantic Analysis (pLSA)

Although block LBP features mentioned above can characterize the pathology images, they are likely to ignore the high-level semantics that may exist in the image. As the high-level semantic model, Bag-of-words (BoW) [12] performs *k*-means clustering on the local feature descriptors to generate the codebook composed of visual words, and then quantizes the descriptors into the words through nearest neighbor principal. Finally the image can be represented by words. However, there are synonyms among visual words, which may cause that the semantics of images are not well revealed. As a well-known topic model, probabilistic latent semantic analysis (pLSA) [9] model aims to describe the image content by the latent topic-level representation learned from the visual words. Moreover, it has simplicity and low computational complexity.

Let **Z** = [$z_1,\ldots,z_T$] be the set of latent topics between documents **D** = [d_1, d_2, …, d_M] and words **W** = [w_1, w_2,…, w_N,]. The goal of pLSA is to learn the latent topic probability distribution through the joint probability distribution of documents **D** and words **W**. Specifically, for image retrieval application, **D** is a dataset of images, and **W** is the collection of visual word representations in the dataset and **Z** can be viewed as the latent variables between **W** and **D**, namely the topic-level representation. The graph model representation of pLSA is shown in Fig. 2.

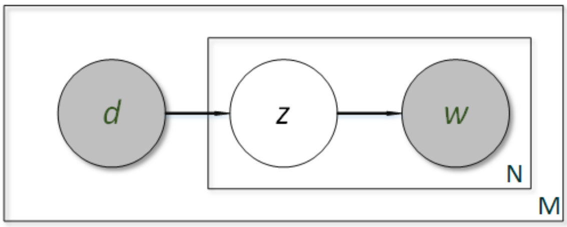

Fig. 2. Graph model representation for pLSA

The joint probability between **W** and **D** is defined by Eq. (2):

$$P(d_i,w_j)=P(d_i)P(w_j\,|d_i),\ P(w_j\,|d_i)=\sum_{k=1}^{T}P(z_k\,|d_i)P(w_j\,|z_k \tag{2}$$

where $P(d_i)$ denotes the probability which d_i occurs, $P(z_k|d_i)$ is probability distribution of latent topic z_k in document d_i and $P(w_j|z_k)$is the probability distribution of topic z_k on word w_j. pLSA model can be viewed as a maximum log-likelihood formulation:

$$\begin{aligned} L &= \sum_{i=1}^{N}\sum_{j=1}^{M} n(d_i,w_j)\log P(d_i,w_j) \\ &= \sum_{i=1}^{N} n(d_i)\left[\log P(d_i)+\sum_{j=1}^{M}\frac{n(d_i,w_j)}{n(d_i)}\log\sum_{k=1}^{T}P(w_j\mid z_k)P(z_k\mid d_i)\right| \end{aligned} \tag{3}$$

where $n(d_i,\ w_j)$ represents the frequency that word w_j occurs in document d_i and $n(d_i)$ denotes the occurrence frequency of d_i. Therefore, the solution of pLSA model is to seek the optimal $P(z_k|d_i)$ and $P(w_j|z_k)$ through expectation-maximization (EM) algorithm [9], and $P(z|d_i)$ is the topic-level representation of the i-th document.

3.2 Low-Rank and Sparse Matrix Decomposition

According to [10], the word-level representation generated by BoW implies both correlated and specific information, and each of these two parts is more robust and discriminative for representing the image content. In this paper, we apply the low-rank and sparse matrix decomposition method to decompose the BoW representation (*i.e.*, the word-level representation) of the images into two parts (*i.e.*, low-rank part and sparse part). After decomposition, the low-rank part can reveal the correlated characteristics of words and the sparse part can indicate the specific characteristics of words. In other words, we obtain two word-level distributions that can describe the generality and specialty of words through low-rank and sparse matrix decomposition.

As mentioned above in Section 3.1, $\mathbf{W} = [w_1, w_2, \ldots, w_N]$ is the collection of BoW representations where w_i is the representation of i-th training image. Thus the decomposition is defined as:

$$\mathbf{W} = \mathbf{L} + \mathbf{S} \tag{4}$$

where **L** and **N** are the low-rank and the sparse matrix. The problem of low-rank and sparse matrix decomposition can be characterized by

$$\begin{gathered} \min_{L,N} \ rank(\mathbf{L}) + \lambda \|\mathbf{S}\|_0 \\ s.t. \mathbf{W} = \mathbf{L} + \mathbf{S} \end{gathered} . \tag{5}$$

The $\| \cdot \|_0$ is zero-norm that counts the non-zero elements in the matrix and $\lambda > 0$ is the coefficient that balances the rank term and the sparsity term. Since the problem is non-convex and hard to solve, it can be approximated by solving (6) according to [13]:

$$\begin{gathered} \min_{L,N} \|\mathbf{L}\|_* + \lambda \|\mathbf{S}\|_1 \\ s.t. \mathbf{W} = \mathbf{L} + \mathbf{S} \end{gathered} \tag{6}$$

The $\| \cdot \|_*$ is the nuclear norm defined as the sum of all singular values. The problem can be solved by the augmented Lagrange multiplier method (ALM) proposed by Lin et al [14].

3.3 Low-Rank and Sparse Matrix Decomposition Based pLSA Model

After the low-rank and sparse matrix decomposition, we obtain a low-rank matrix **L** which can characterize the correlated part of words and a sparse matrix **N** which can characterize the specific part of words. Each column vector l_i of the matrix $\mathbf{L} = [l_1, l_2, \ldots, l_N]$ can be regarded as the representation of correlated characteristics of the i-th training image, and each column vector n_i of the matrix $\mathbf{S} = [s_1, s_2, \ldots, s_N]$ implies the specific characteristics. Therefore, instead of w_i, we respectively apply l_i and n_i for the word-level representations of i-th image, and then use them to train two pLSA models. Note that l_i and n_i are L_1-normalized after absolute operation.

The flow chart of our work is presented in Fig. 3. First we extract the 59-dimensional block LBP histogram for each pathology images in the training set. Then the codebook can be gained through k-means and the word-level representation (namely w_i in matrix **W**) corresponding to each image is quantized. After the low-rank and sparse matrix decomposition step, the matrices **L** and **S** take place of **W** to be the word-level representations. EM algorithm is respectively used to compute the optimal P(z|d) and P(w|z) of these two representations, and the combination of P(z|d) is the final topic-level representation of each image. In the test stage, the input ROI will be converted to the topic-level representation learned from correlated and specific word-level distributions. After computing the similarities between ROI image and the images stored in the database, the top R similar images along with the confirmed diagnosis information are returned to the CAD system. By comparing ROI with these returned images, pathologists can make a more reliable diagnosis decision.

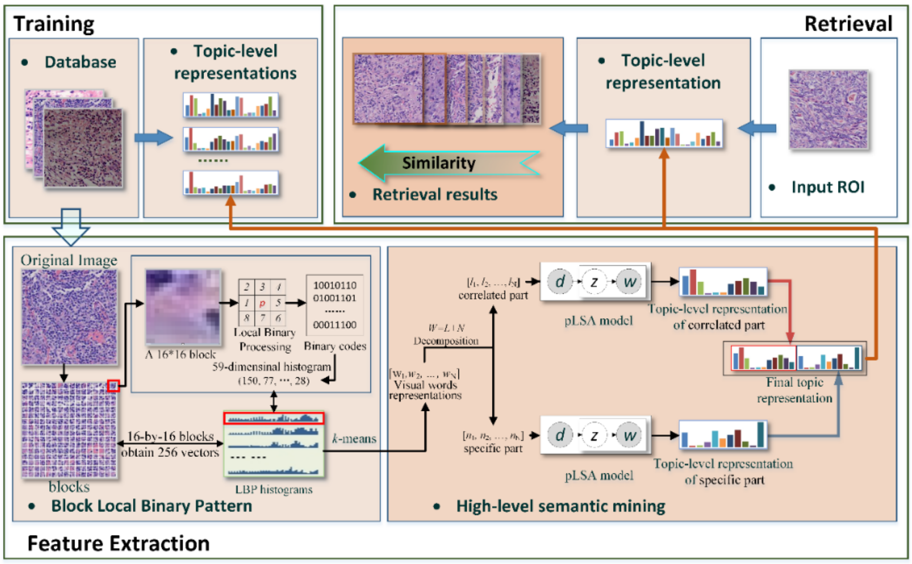
Training
Database
Topic-level representations
Retrieval
Topic-level representation
Input ROI
Similarity
Retrieval results
Original Image
A 16*16 block
Local Binary Processing
Binary codes
10010110
01001101
00011100
59-dimensinal histogram
(150, 77, ···, 28)
blocks
16-by-16 blocks obtain 256 vectors
LBP histograms
k-means
Block Local Binary Pattern
$[l_1, l_2, ..., l_M]$ correlated part
$W=L+N$ Decomposition
$[w_1, w_2, ..., w_N]$ Visual words representations
$[n_1, n_2, ..., n_K]$ specific part
d
z
w
pLSA model
Topic-level representation of correlated part
Topic-level representation of specific part
Final topic representation
High-level semantic mining
Feature Extraction

Fig. 3. The flow chart of our retrieval framework

4 Experiment

Our proposed method is evaluated on the pathology image database for breast cancer with confirmed diagnosis information, which is from Motic digital slide database for the yellow race [20]. The image database consists of 5 categories and 600 images (256×256, 20x magnification) for each category, as shown in Fig. 4.

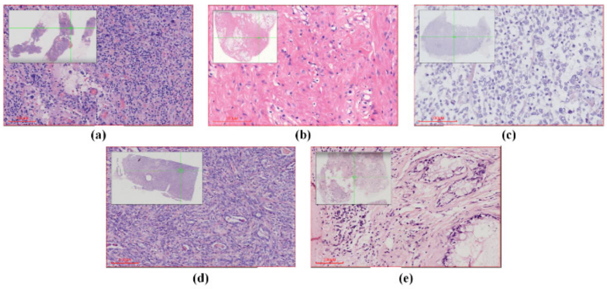

Fig. 4. Five categories of digital pathology slides. (a) Basal-like carcinoma (BLC). (b) Breast myofibroblastoma (BMFB). (c) Invasive breast cancer (IBC). (d) Low-grade adenosquamous carcinoma (LGASC). (e) Mucinous cystadenocarcinoma (MCA).

To evaluate the performance of our method, we compare it with block-LBP-based BoW (LPB-BoW), block-LBP-based pLSA (LBP- pLSA), along with approaches of Caicedo et al. [6] and Kowal et al [7]. Note that we perform 20 times to randomly select 300 images of each category for training and the remaining for test. For each time, we calculate the mean Average Precision (mAP) these five methods for top 20 returned images through cosine distance based similarity measure. Table 1 shows the performances of these five methods.

Table 1. Performance comparison at the top 20 returns of five methods

Algorithm	Performance
Kowal et al. [7]	53.1±1.61
Caicedo et al. [6]	70.2±1.06
LPB-BoW	74.4±1.17
LBP-pLSA	76.3±0.41
Our method	78.6±0.49

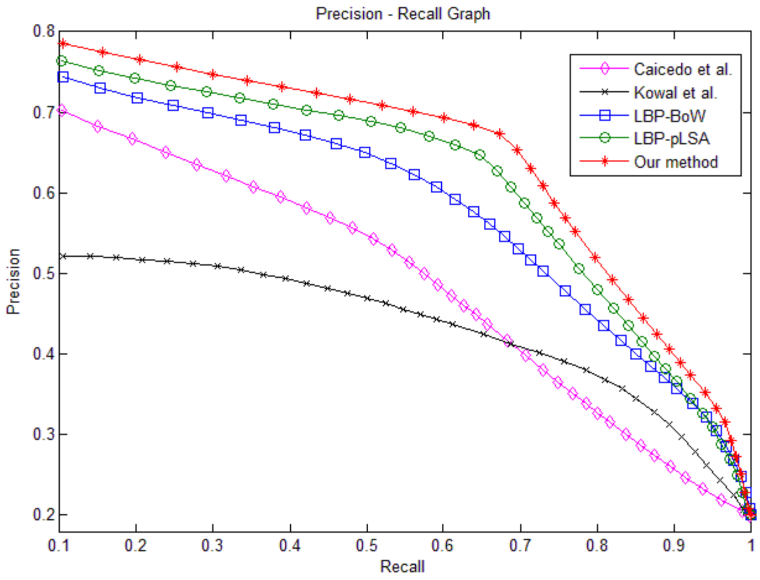

Fig. 5. Precision-recall curves of five methods

As can be seen from Table 1, compared with the methods proposed by Caicedo et al. [6] and Kowal et al [7], LPB-BoW has superior retrieval performance. It may be due to the fact that block-LBP features can effectively describe the spatial property of texture structure, and in the other hand, it may benefit from the semantic characterization ability of BoW. Particularly, as pLSA model overcomes the limitation of BoW, LBP-pLSA and our method are better. It should be noted that our method is more excellent than LBP-pLSA, since it can discover the correlated and specific parts of the visual word distribution which leads to the more discriminant word-level representation. The precision-recall curves of these methods are presented in Fig. 5. It indicates that our method overall outperforms the others.

5 Conclusion

In this paper, we propose a novel pathology image retrieval method for breast cancer. Block LBP descriptor is used to describe the spatial characteristics of texture structure. Then they are applied to generate into visual word representation by BoW scheme. After low-rank and sparse composition operating, the word-level representation of each image is decomposed into correlated part and specific part. Based on these two parts, two pLSA models are leant to mine the high-level semantics existed in the images. Finally, each image is represented by the combined topics of the two pLSA models. Experiments on the pathology image database for breast cancer demonstrate the effectiveness of our method. Further research will aim to apply Local Sensitive Hashing (LSH) to boost the efficiency of retrieval when facing large database.

Acknowledgement. This work was supported by the National Natural Science Foundation of China (No. 61071137), and the 973 Program of China (Project No. 2010CB327900).

References

1. Rebecca, S., Deepa, N., Ahmedin, J.: Cancer Statistics. CA Cancer Journal for Clinicians 63(1), 11–30 (2013)
2. Li, N., Zheng, R.S., Zhang, S.W., Zou, X.N., Zeng, H.M., Dai, Z., Chen, W.Q.: Analysis and Prediction of Breast Cancer Incidence Trend in China. Chinese Journal of Preventive Medicine 46(8), 703–707 (2012)
3. Fu, X.L.: The Atlas for Pathologic Diagnosis of Breast Tumours. Scientifics and Technical Documents Publishing House, Beijing (2013)
4. Xue, Z.Y., Long, L.R., Antani, S., Thoma, G.R.: Pathological-based Vertebral Image Retrieval. In: IEEE International Symposium on Biomedical Imaging: From Nano to Macro (ISBI), Chicago, pp. 1893–1896 (2011)
5. Lijia, Z., Shaomin, Z., Dazhe, Z., Hong, Z., Shukuan, L.: Medical Image Retrieval Using Sift Feature. In: IEEE 2nd International Congress on Image and Signal Processing, pp. 1–4. Tianjin (2009)
6. Caicedo, J.C., Izquierdo, E.: Combining Low-level Features for Improved Classification and Retrieval of Histology Images. Transactions on Mass-Data Analysis of Images and Signals 2(1), 68–82 (2010)
7. Marek, K., Paweł, F., Andrzej, O., Józef, K., Roman, M.: Computer-aided Diagnosis of Breast Cancer Based on Fine Beedle Biopsy Microscopic Images. Computers in Biology and Medicine 43(10), 1563–1572 (2013)
8. Ojala, T., Pietikäinen, M., Harwood, D.: A Comparative Study of Texture Measures with Classification Based on Featured Distributions. Pattern Recognition 29(1), 51–59 (1996)

9. Hofmann, T.: Probabilistic Latent Semantic Analysis. In: The 22nd Annual ACM Conference on Research and Development in Information Retrieval, San Francisco, pp. 289–296 (1999)
10. Zhang, C., Liu, J., Tian, Q., et al.: Image Classification by Non-Negative Sparse Coding, Low-Rank and Sparse Decomposition. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1673–1680 (2011)
11. Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(7), 971–987 (2002)
12. Fei-Fei, L., Perona, P.: A Bayesian Hierarchical Model for Learning Natural Scene Categories. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), San Diego, pp. 524–531 (2005)
13. Candes, E., Li, X., Ma, Y., Wright, J.: Robust Principal Component Analysis? Journal of the ACM 58(3), 11 (2011) (submitted)
14. Lin, Z., Chen, M., Ma, Y.: The Augmented Lagrange Multiplier Method for Exact Recovery of Corrupted Low-Rank Matrices. arXiv preprint arXiv, 1009.5055 (2010)
15. Motic digital slide database, `http://med.motic.com/SlideLibraryList.aspx`