Mining Histopathological Images via Composite Hashing and Online Learning

Xiaofan Zhang[1], Lin Yang[2], Wei Liu[3], Hai Su[2], and Shaoting Zhang[1]

[1] Department of Computer Science, UNC Charlotte, NC, USA
[2] Department of Biostatistics, University of Kentucky, Lexington, KY, USA
[3] IBM T. J. Watson Research Center, Yorktown Heights, NY, USA

Abstract. With a continuous growing amount of annotated histopathological images, large-scale and data-driven methods potentially provide the promise of bridging the semantic gap between these images and their diagnoses. The purpose of this paper is to increase the scale at which automated systems can entail scalable analysis of histopathological images in massive databases. Specifically, we propose a principled framework to unify hashing-based image retrieval and supervised learning. Concretely, composite hashing is designed to simultaneously fuse and compress multiple high-dimensional image features into tens of binary hash bits, enabling scalable image retrieval with a very low computational cost. Upon a local data subset that retains the retrieved images, supervised learning methods are applied on-the-fly to model image structures for accurate classification. Our framework is validated thoroughly on 1120 lung microscopic tissue images by differentiating adenocarcinoma and squamous carcinoma. The average accuracy is 87.5% with only 17ms running time, which compares favorably with other commonly used methods.

1 Introduction

Lung cancer is one of the most common and deadly cancer in the world. There were an estimated 228,190 new cases and 159,480 deaths in 2013 [12]. Four typical histologic types of lung cancers include adenocarcinoma, squamous carcinoma, small cell carcinoma, and large cell carcinoma with the 5-year survival rate below 20%, each of which requires a different treatment [5]. Consequently, early diagnosing and differentiating these histologic types is very important. Bronchial biopsy is one of the most effective diagnosis methods for lung cancer. Traditional manual exams of microscopy are labor intensive, time consuming, and not too accurate [2] as well. Therefore, Computer Aided Diagnosis (CAD) systems have been investigated to automatically analyze histopathological images of lung cancer. For examples, Kayser *et al.* [6] employed textural and morphometric features including image entropy and construction of skeleton to classify tumor and non-tumor lung diseases. The validity of quantitative nuclear image features is evaluated in discriminating small cell and non-small cell lung cancers [15]. Mijović *et al.* [11] proposed a model to classify histopathological images by manually editing the binary images, extracting seven nuclear variables (*e.g.*, nuclear area, equivalent diameter, volume of equivalent sphere, perimeter, etc.), and analyzing the statistics with stepwise linear discriminant analysis and the Mann-Whitney test. Promising experimental results are shown in classifying small cell and non-small cell lung cancers.

P. Golland et al. (Eds.): MICCAI 2014, Part II, LNCS 8674, pp. 479–486, 2014.

Most existing systems have focused on the classification of small cell vs. non-small cell (*i.e.*, adenocarcinoma, squamous carcinoma, and large cell carcinoma) lung cancers, since the treatment plan is mainly decided by this result. However, it is also clinically important to differentiate the adenocarcinoma and squamous carcinoma, both of which belong to non-small cell lung cancer, because the management protocols might differ in different tumour types [4]. Unfortunately, there are few attempts in dealing with this problem, due to the subtle differences of their histopathological images which make a large obstacle for classification [11]. In addition, most systems in this area can only handle relatively small datasets. The development of scalable image analysis algorithms does not keep pace with the increasing quantity of histopathological images. Hence, there is an urgent need to build a more innovative framework to entail efficient analysis in massive-scale databases.

In this paper, we tackle the challenging scalability issue and propose a scalable framework for the early diagnosis of two non-small cell types, *i.e.*, adenocarcinoma and squamous carcinoma. Specifically, we design an effective framework to couple scalable content-based image retrieval (CBIR) and online learning via a "coarse-to-fine" scheme. In contrast to traditional classification methods that directly compute the likelihood of the diagnostic results, scalable CBIR approaches open a new avenue for analyzing images in potentially massive databases. We develop a Composite Anchor Graph Hashing algorithm to simultaneously fuse multiple high-dimensional image features and compress them into tens of binary hash bits, thereby enabling rapid retrieval (in sublinear or even constant time) of relevant instances among large database. Such retrieved results can be utilized to support clinical decisions or indicate the most likely diagnosis using the majority logic. After discovering the subset of images, we are able to apply data-driven supervised learning methods that are specific to the input image (*i.e.*, queries) on-the-fly to achieve accurate classification. We conduct extensive experiments on thousands of lung microscopic tissue images for the differentiation of adenocarcinoma and squamous carcinoma. The experimental results demonstrate high classification accuracy and favorable computational efficiency of our proposed framework.

2 Methodology

2.1 Overview

Fig. 1 shows two main phases of our proposed framework. The offline phase compresses texture features into short binary codes, which enables real-time search even among large-scale databases and high-dimensional features. Specifically, Difference of Gaussians is used to detect interest points, as shown in Fig. 2. These points approximately indicate the locations of nuclei. Then, two types of texture features [9,3] are extracted at each point. In fact, this setting of detections and features has been widely used in both general image analysis and histopathological images [1]. In this specific problem, it is desired to maximally utilize these millions of cell-level features (thousands in each image). Therefore, we quantize them into high-dimensional feature vectors, *i.e.*, 10,000

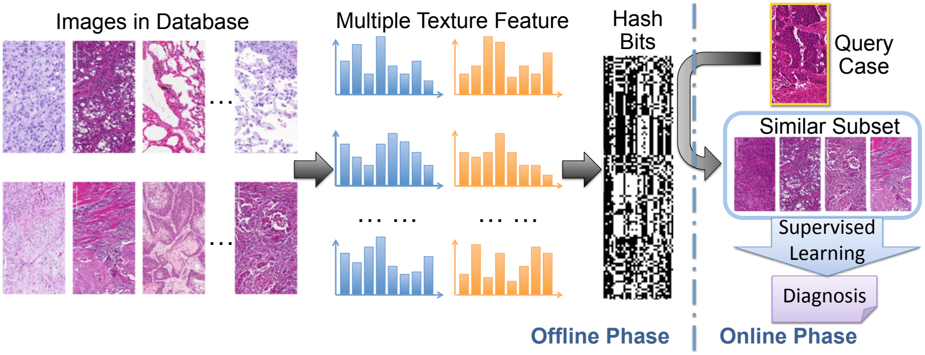

Fig. 1. Overview of our proposed framework for scalable analysis of histopathological images

length, via bag-of-words [13]. However, using features with such large size may adversely affect the computational efficiency. Therefore, we propose to simultaneously fuse and compress them into tens of bits using Composite Anchor Graph Hashing algorithm.

During the online phase, we extract same types of texture features from the input image (*i.e.*, query) and compress them as binary bits using the same algorithm. By comparing this with the binary bits extracted offline, we can efficiently retrieve images with morphological profiles most relevant and consistent to the query, and the retrieved images indicate the most likely diagnosis. In fact, this is an example of scalable content-based image retrieval (CBIR). However, despite the efficiency and scalability of our composite hashing method, it may result in an unordered set where exact or near-exact matches may be obscured within a large database due to noisy features or similar instances. Therefore, we further propose to employ supervised learning methods on-the-fly to build a finer model in the retrieved set for accurate classification and diagnosis.

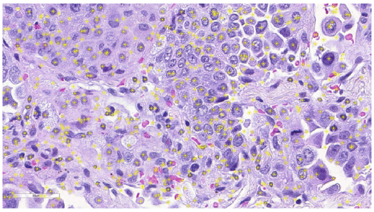

Fig. 2. Detected interest points (marked as yellow stars)

2.2 Hashing via Composite Graphs

For scalable analysis of histopathological images, we develop an efficient algorithm to search among the high-dimensional and large-scale feature space, *i.e.*, locate visually-similar images in real-time. Specifically, hashing algorithm is employed to map the neighbors in the high-dimensional feature space into similar binary codes in the Hamming space.

Notations and Problem Definition: Denote $\boldsymbol{x} \in \mathbb{R}^{n \times d}$ as the d-dimensional texture features of n histopathological images, Y as the r bits binary code compressed from $\boldsymbol{x}$, and A as the $n \times n$ similarity matrix in which A_{ij} means the similarity between features ($\boldsymbol{x}_i$ and $\boldsymbol{x}_j$). The objective function of such mapping is usually formulated as:

$$\begin{aligned} \min_{Y} \quad & \frac{1}{2} \sum_{i,j=1}^{n} \|Y_i - Y_j\|^2 A_{ij} = tr(Y^\top L Y) \\ s.t. \quad & Y \in \{1, -1\}^{n \times r},\ \mathbf{1}^\top Y = 0,\ Y^\top Y = n I_{r \times r} \end{aligned} \tag{1}$$

where L is the graph Laplacian that is defined as $L = D - A$, and $D = diag(A\mathbf{1})$ with $\mathbf{1} = [1, \cdots, 1]^\top \in \mathbb{R}^n$. To solve this NP-hard problem, we can apply spectral relaxation to drop the integer constraint and allow $Y \in \mathbb{R}^{n \times r}$.

Composite Graphs: An important module of this hashing formulation is to compute the graph Laplacian, *i.e.*, building the underlying nearest neighbor graph. Traditional methods are not computationally efficient, especially for a large number of high-dimensional features. In fact, its computational complexity is $\mathcal{O}(dn^2)$, which is intractable for large n and d. Recently proposed Anchor Graph Hashing [8] is able to alleviate this computational bottleneck by approximating the data neighborhood structure with a small set of "anchors". However, same as most previous hashing algorithms, it focuses on modeling and compressing a single type of feature, while it is necessary to fuse multiple features in histopathological image analysis to achieve a better accuracy of diagnosis. Therefore, we propose Composite Anchor Graph to efficiently fuse multiple features and build a unified graph. Specifically, m anchors ($\mathcal{U} = \{\boldsymbol{u}_j \in \mathbb{R}^d\}_{j=1}^{m}$ $(m \ll n)$) are chosen to construct the graph:

$$Z_{i,j} = \begin{cases} \frac{exp(-\mathcal{D}^2(\boldsymbol{x}_i, \boldsymbol{u}_j)/t)}{\sum_{j' \in \langle i \rangle} exp(-\mathcal{D}^2(\boldsymbol{x}_i, \boldsymbol{u}_{j'})/t)}, & \forall j \in \langle i \rangle \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

where Z is a highly sparse $n \times m$ matrix that represents truncated similarities between all n data points and m, t denotes the bandwidth parameter anchors, $\langle i \rangle \subset [1 : m]$ is the indices of s $(s \ll m)$ nearest anchors of point $\boldsymbol{x}_i$ in $\mathcal{U}$ according to a distance function $\mathcal{D}()$. Compared with constructing a exact kNN graph with a quadratic cost $\mathcal{O}(dn^2)$, the cost for constructing an Anchor Graph is linear to n when m is small, which is significantly faster.

Based on this efficient graph construction, we fuse multiple types of features on the distance level, by normalizing and averaging them with the assumption that each individual feature contributes equally [10] or using weights learned from training data [18]. This scheme naturally fuses multiple features in the anchor graph and hence approximates the similarity matrix A by $\hat{A} = Z\Lambda^{-1}Z^\top$, where $\Lambda = diag(Z^\top \mathbf{1}) \in \mathbb{R}^{m \times m}$. Subsequently, the graph Laplacian is represented as $L = I - \hat{A}$. The eigenvectors of $\hat{A}$ can be easily solved by utilizing its low-rank property. The results of $m \times m$ matrix $\Lambda^{-1/2}Z^\top Z\Lambda^{-1/2}$ are eigenvector-eigenvalue pairs $\{(\boldsymbol{v}_k, \sigma_k)\}_{k=1}^{r}$, where $1 > \sigma_1 \geq \cdots \geq \sigma_r > 0$. Denote V as $[\boldsymbol{v}_1, \cdots, \boldsymbol{v}_r] \in \mathbb{R}^{m \times r}$, Σ as $diag(\sigma_1, \cdots, \sigma_r) \in \mathbb{R}^{r \times r}$, W as $\sqrt{n}\Lambda^{1/2}V\Sigma^{-1/2}$. Therefore, the spectral embedding matrix Y is computed as:

$$Y = \sqrt{n} Z \Lambda^{-1/2} V \Sigma^{-1/2} = ZW \tag{3}$$

Hash Function: Using the Laplacian of this Composite Anchor Graph, a general hash function $h_k(\boldsymbol{x}) = sgn(\phi_k(\boldsymbol{x}))$ can be obtained by generalizing its eigenvectors to the eigenfunctions with Nyström method [16].

$$h_k(\boldsymbol{x}) = sgn(\boldsymbol{w}_k^\top \boldsymbol{z}(\boldsymbol{x})),\ k = 1, \cdots, r \tag{4}$$

where $\boldsymbol{w}_k = \sqrt{\frac{n}{\sigma_k}} \Lambda^{-1/2} \boldsymbol{v}_k,\ \boldsymbol{z}(\boldsymbol{x}) = \frac{\left[\delta_1 exp(-\mathcal{D}^2(\boldsymbol{x},\boldsymbol{u}_1)/t), \cdots, \delta_m exp(-\mathcal{D}^2(\boldsymbol{x},\boldsymbol{u}_m)/t)\right]^\top}{\sum_{j=1}^m \delta_j exp(-\mathcal{D}^2(\boldsymbol{x},\boldsymbol{u}_j)/t)}$, $\delta_j \in \{1, 0\}$ and $\delta = 1$ if and only if anchor $\boldsymbol{u}_j$ is one of s nearest anchors of sample $\boldsymbol{x}$ in $\mathcal{U}$.

Consequently, hashing with Composite Anchor Graphs can be formulated as first calculating the sparse similarity matrix Z between fused texture features and anchors, and then projecting each Z_i onto the vectors in W.

The computational complexity of this module is $O(dmn + m^2n + (s+1)rn)$, where $O(dmn)$ is the time to build the composite anchor graph, $O(m^2n + srn)$ is to solve r graph Laplacian eigenvectors retained in the spectral embedding matrix Y, and $O(rn)$ represents the time for compressing Y into binary code. The query time is sublinear or even constant when using a hash table. Therefore, the resulting tens of bits significantly improve the computational efficiency of image retrieval, without sacrificing the accuracy.

2.3 Learning on Retrieved Images

Our composite hashing method is able to effectively retrieve a set of visually similar images via simultaneously unifying and compressing multiple features. These retrieved images can be directly used for decision support or diagnosis using the majority logic. Despite the efficacy of hashing methods and their successful uses in medical image analysis [19,7], this scheme still may not be accurate enough since they tend to produce unordered results, *i.e.*, same hamming distance or hash entries for different images. In addition, the "semantic gap" problem also exists in our hashing method as there is no association between the texture features and the diagnosis information. Supervised learning methods such as Support Vector Machine (SVM) may alleviate this semantic gap problem, but they may not be able to learn an effective model over the whole dataset due to the complexity of histopathological images. Furthermore, the computational complexity and scalability is also an issue.

Since our hashing method is able to locate a small number of images with negligible cost in running time, we propose to apply supervised learning methods (SVM in our system) on-the-fly among the local set of images. This "coarse-to-fine" and "online" scheme is able to build local models adaptively and accurately, while still maintaining the computational efficiency. In addition, this model can incorporate new samples online without any retraining, as supervised learning is only performed locally during runtime. In the experiments, we demonstrate that our framework significantly outperforms traditional hashing algorithms or supervised learning in the global-level.

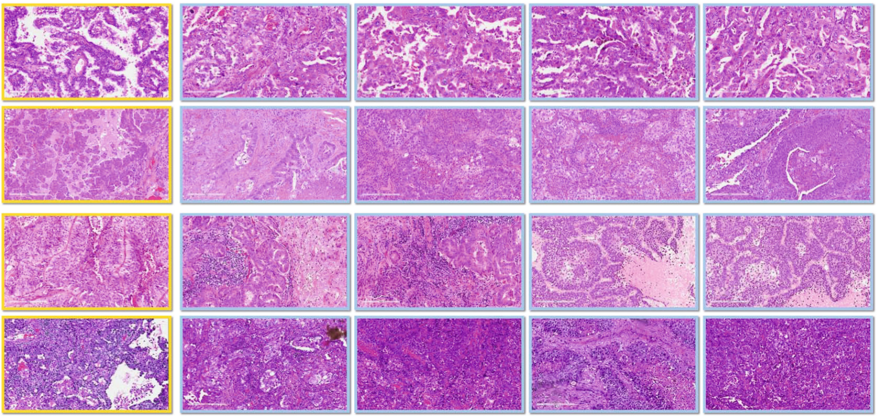

Fig. 3. Examples in retrieval. Query marked in yellow, and retrieved images marked in blue. The first two rows are adenocarcinoma, and the last two rows are squamous carcinoma.

3 Experiments

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

Experimental Setting: 1120 images ($2K \times 1K$ pixels each) are sampled from 112 patients (57 adenocarcinoma and 55 squamous carcinoma). Leave-one-patient-out validation is used to evaluate the accuracy of classification. All the experiments are conducted on a 3.40GHz CPU with 4 cores and 16G RAM, in a MATLAB implementation. Around 3000 keypoints are detected in each image, which result in millions of texture descriptors in the whole dataset. To effectively leverage such massive information, we use high-dimensional features. Specifically, we extract widely-used texture descriptors [9,3] from each image and quantized each type of descriptors with 10,000 dimensional bag-of-words. We validate the efficacy of our Composite Anchor Graph by fusing different types of descriptors, and also compare the proposed framework with the classical classifiers such as SVM and k-nearest neighbors (kNN), which have been used for histopathological image analysis [17,14].

Evaluation of Composite Hashing: We validate our composite hashing by comparing the average accuracy of differentiating adenocarcinoma and squamous carcinoma using single feature or fused features. Concatenating feature vectors with proper normalization is chosen as the baseline as this is the most-widely used approach for feature fusion. However, this scheme does not consider the preservation of local graph structures and hence only marginally improves each feature by less than 1%, *i.e.*, 73.4%, 72.6% and 74.1% for two types of features and the fused one, respectively. In contrast, our composition scheme considers such local manifold information and fuse features in the distance level. Therefore, it achieves a higher accuracy (76.8%). In addition, we also show the retrieved images using our composite hashing algorithm in Fig. 3. In fact, this CBIR task is very challenging since the visual dissimilarity of certain images in different categories is quite subtle. Our accurate results demonstrate the efficacy of the composite hashing and the texture features. These retrieved results are clinically relevant to the query and thus potentially useful for decision support.

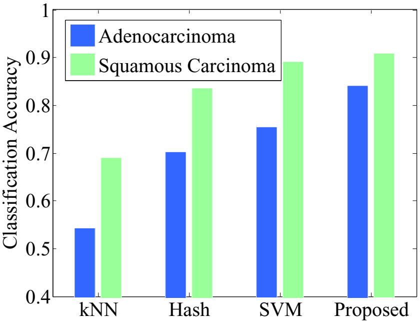

Fig. 4. Comparison of the accuracy using different classification models

Evaluation of the Framework: Fig. 4 shows the comparison of different classification methods, measured by the diagnosis accuracy of the adenocarcinoma and squamous carcinoma. kNN is an effective approach for classification and CBIR, and always serves as the baseline. Here we use Euclidean distance and set the k as 5. However, it is not able to achieve high accuracy in this challenging problem (61.6% in average), due to the large variations of the staining and cell distribution. In addition, it is not computationally efficient and hence adversely affect its scalability, *i.e.*, linear to the number of images and feature dimensions. Our composite hashing method significantly improves the efficiency, *i.e.*, sub-linear when using hamming distance or even constant time when using hash table. In our experiments, it merely takes 2ms for each query with 100 anchors and 36 bits, which is around 50 times faster than kNN. Furthermore, these compressed features even achieve higher accuracy (76.8% in average), owing to the preservation of local structures (*i.e.*, manifold) when constructing anchor graphs, and the effective feature fusion scheme. Not surprisingly, SVM with RBF kernel outperforms both kNN and hashing, as it leverages the supervised information to bridge the semantic gap. The average accuracy of SVM is 82.1%. However, learning a classification hyper plane on the whole dataset results in a highly unbalanced result, *i.e.*, 75.4% for adenocarcinoma and 89.0% for squamous carcinoma. Furthermore, SVM requires offline training upon the whole dataset, which sacrifices the scalability for appending new data. Our proposed framework achieves the best accuracy, with 87.5% in average, 84.2% for adenocarcinoma, and 90.9% for squamous carcinoma. This is substantially better than other methods, and is also much more balanced than SVM. This improvement demonstrates the efficacy of applying supervised learning method on the local set of images discovered by our hashing method. The additional computational complexity is from the learning method applied on-the-fly, which takes 15ms for each image and is independent of the database's size. Therefore, the running time of our proposed framework is 17ms, and there is no extra cost in the asymptotic analysis when expanding the size of database, which is a significant benefit compared to traditional methods.

4 Conclusion

In this paper, we proposed a scalable framework for histopathological image analysis, and validated the efficacy of our framework on thousands of lung microscopic tissue images. The main contribution lies in unifying scalable CBIR based on hashing and supervised learning methods on-the-fly, which turns out to exhibit high classification accuracy and computational efficiency. Furthermore, the Composite Anchor Graph Hashing algorithm was developed to improve hashing quality by simultaneously fusing and compressing multiple high-dimensional image features. In the future, we intend to investigate visual analytical methods for the intelligent and interactive visualization of histopathological images, based on the outcome of this paper.

References

1. Caicedo, J.C., Cruz, A., Gonzalez, F.A.: Histopathology image classification using bag of features and kernel functions. In: Combi, C., Shahar, Y., Abu-Hanna, A. (eds.) AIME 2009. LNCS, vol. 5651, pp. 126–135. Springer, Heidelberg (2009)
2. Cataluña, J.J.S., Perpiñá, M., Greses, J.V., Calvo, V., Padilla, J.D., París, F.: Cell type accuracy of bronchial biopsy specimens in primary lung cancer. CHEST 109(5) (1996)
3. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR, vol. 1, pp. 886–893 (2005)
4. Edwards, S., Roberts, C., McKean, M., Cockburn, J., Jeffrey, R., Kerr, K.: Preoperative histological classification of primary lung cancer: accuracy of diagnosis and use of the non-small cell category. Journal of Clinical Pathology 53(7), 537–540 (2000)
5. Freeman, D.L.: Harrison's principles of internal medicine. JAMA: The Journal of the American Medical Association 286(8), 506 (2001)
6. Kayser, G., Riede, U., Werner, M., Hufnagl, P., Kayser, K.: Towards an automated morphological classification of histological images of common lung carcinomas. Elec. J. Pathol. Histol. 8, 022–03 (2002)
7. Liu, J., Zhang, S., Liu, W., Zhang, X., Metaxas, D.: Scalable mammogram retrieval using anchor graph hashing. In: ISBI (2014)
8. Liu, W., Wang, J., Kumar, S., Chang, S.F.: Hashing with graphs. In: ICML, pp. 1–8 (2011)
9. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. IJCV 60(2), 91–110 (2004)
10. Makadia, A., Pavlovic, V., Kumar, S.: Baselines for image annotation. IJCV 90(1), 88–105 (2010)
11. Mijović, Ž., Mihailović, D., Kostov, M.: Discriminant analysis of nuclear image variables in lung carcinoma. Facta Universitatis-Series: Medicine and Biology 15(1), 28–32 (2008)
12. Siegel, R., Naishadham, D., Jemal, A.: Cancer statistics. CAJC 63(1), 11–30 (2013)
13. Sivic, J., Zisserman, A.: Video google: A text retrieval approach to object matching in videos. In: ICCV, pp. 1470–1477 (2003)
14. Tabesh, A., Teverovskiy, M., Pang, H.Y., Kumar, V.P., Verbel, D., Kotsianti, A., Saidi, O.: Multifeature prostate cancer diagnosis and gleason grading of histological images. TMI 26(10), 1366–1378 (2007)
15. Thunnissen, F., Diegenbach, P., Van Hattum, A., Tolboom, J., van der Sluis, D., Schaafsma, W., Houthoff, H., Baak, J.R.: Further evaluation of quantitative nuclear image features for classification of lung carcinomas. Pathology-Research and Practice 188(4), 531–535 (1992)
16. Williams, C., Seeger, M.: Using the nyström method to speed up kernel machines. In: NIPS (2001)
17. Yang, L., Chen, W., Meer, P., Salaru, G., Goodell, L.A., Berstis, V., Foran, D.J.: Virtual microscopy and grid-enabled decision support for large-scale analysis of imaged pathology specimens. TITB 13(4), 636–644 (2009)
18. Zhang, S., Huang, J., Huang, Y., Yu, Y., Li, H., Metaxas, D.N.: Automatic image annotation using group sparsity. In: CVPR, pp. 3312–3319 (2010)
19. Zhang, X., Liu, W., Zhang, S.: Mining histopathological images via hashing-based scalable image retrieval. In: ISBI (2014)