Fine-Grained Histopathological Image Analysis via Robust Segmentation and Large-Scale Retrieval

Xiaofan Zhang[1], Hai Su[2], Lin Yang[2], Shaoting Zhang[1*]
[1]University of North Carolina at Charlotte, Charlotte, NC, 28223, USA.
[2]University of Florida, Gainesville, FL, 32611, USA.
*Corresponding author: shaoting@cs.rutgers.edu

Abstract

Computer-aided diagnosis of medical images requires thorough analysis of image details. For example, examining all cells enables fine-grained categorization of histopathological images. Traditional computational methods may have efficiency issues when performing such detailed analysis. In this paper, we propose a robust and scalable solution to achieve this. Specifically, a robust segmentation method is developed to delineate region-of-interests (e.g., cells) accurately, using hierarchical voting and repulsive active contour. A hashing-based large-scale retrieval approach is also designed to examine and classify them by comparing with a massive training database. We evaluate this proposed framework on a challenging and important clinical use case, i.e., differentiation of two types of lung cancers (the adenocarcinoma and the squamous carcinoma), using thousands of histopathological images extracted from hundreds of patients. Our method has achieved promising performance, i.e., 87.3% *accuracy and* 1.68 *seconds by searching among half-million cells.*

1. Introduction

Recently, digitized tissue histopathology for microscopic examination and automatic disease grading has become amenable to the application of computerized image analysis and computer-aided diagnosis [14]. Many methods have been proposed to tackle this important and challenging use case, by investigating object level and spatially related features [2, 31] and employing different classification schemes [28, 26, 6]. Despite rapid progress in recent years, the main challenge in terms of the computational techniques is the need of analyzing all individual cells for accurate diagnosis, since the differentiation of most disease grades highly depends on the cell-level information, such as its morphology, shape and appearance. In fact, most use cases of medical image analysis require such thorough examination. However, this is very time-consuming. For example, a whole slide histopathological image may have billions of pixels, and even a region-of-interest (ROI) image contains thousands of cells. Analyzing each cell is computational inefficient using traditional methods, if not infeasible. Therefore, many previous methods encode the whole image as holistic features by representing the statistics of cell-level information (e.g., architecture features [6] and frequency of local textures [33]), and may compress high-dimensional features to improve the computational efficiency [33, 34]. Despite the compactness and hence the efficiency, information loss is inevitable in such holistic representation.

In this paper, we design an automatic framework to conduct large-scale analysis of histopathological images, which can examine millions of cells efficiently. Our solution includes two important modules, robust segmentation and large-scale retrieval. Given a new image to be diagnosed, our system automatically segments all cells and efficiently discovers the most relevant cells by comparing them with a large-scale training database (e.g., millions of cells extracted from thousands of images). The diagnosis is decided by classifying each cell using the majority logic. We conduct extensive experiments to differentiate lung cancers, a very challenging use case of histopathological image analysis, using a large dataset containing thousands of lung microscopic tissue images acquired from hundreds of patients. Early diagnosis and differentiation of lung cancers is clinically important due to their different management protocols [8, 10]. However, few efforts have been put on the classification of the adenocarcinoma and squamous carcinoma, i.e., two types of lung cancers, which need cell-level examination to differentiate. Our proposed framework effectively solves this challenging problem, achieving 87.3% accuracy by searching a massive database of half million cells extracted from this dataset, compared favorably with other popular methods for histopathological image analysis.

The major contribution of this paper is twofold. 1) A comprehensive framework is designed to analyze histopathological images by examining all cells. This framework opens a new avenue for the fine-grained classification using large-scale databases of cell images. 2) We propose a carefully designed learning method that assigns probabilistic-based importance to different hash values or entries. This scheme alleviates several intrinsic problems of using traditional hashing methods for classification, and significantly improves the accuracy.

The rest of the paper is organized as follows. Section 2 reviews relevant work of content-based image retrieval for medical image analysis. Section 3 presents the framework of our proposed method for realtime cell classification via hashing. Section 4 shows the experimental results on lung microscopic tissue images. Concluding remarks are given in Section 5.

2. Related Work

Content-Based Image Retrieval (CBIR) is an effective approach in analyzing medical images. It supports doctors to make clinical decisions by retrieving and visualizing relevant medical images with diagnosis information. To this end, many systems and methods have been developed. For examples, Comaniciu et al. [4] designed a CBIR system to support decision making in clinical pathology. In this system, fast color segmenter is used to extract cell features including shape, area, and texture of the nucleus. Its performance was compared with that of a human expert on a database containing two hundred digitized specimens. Dy et al. [7] described a new hierarchical approach of CBIR based on multiple feature sets and a two-step approach. The query image is classified into different classes with discriminative features, and similar images are searched in the predicted class with the features customized to differentiate subclasses. Greenspan et al. [13] proposed a CBIR system that consists of a continuous and probabilistic image-representation scheme. It uses Gaussian mixture models (GMM) and information-theoretic image matching via the Kullback-Leibler (K-L) measure to match and categorize X-ray images by body regions. Song et al. [27] designed a hierarchical spatial matching-based image retrieval method using spatial pyramid matching to extract and represent the spatial context of pathological tissues effectively. Recently, Foran et al. [9] designed a CBIR system to analyze tissue microarrays by harnessing the benefits of high-performance computing and grid technology.

One of the main limitations of these systems is the scalability. To analyze large-scale datasets, one needs to design efficient CBIR methods. With the goal of comparing CBIR methods on a larger scale, ImageCLEF and VISCERAL provide benchmarks for medical image retrieval tasks [23, 18]. In our use case, it is necessary to retrieve among million instances in real-time to conduct cell-level analysis of histopathological images. To this end, hashing-based methods have been investigated, which enable fast approximated nearest neighbors (ANN) search to deal with the scalability issue [5, 30, 17, 21, 11, 19]. Recent representative methods include, but are not limited to, kernelized locality-sensitive hashing [17], weakly-supervised hashing in kernel space [22], semi-supervised hashing [29] and supervised hashing [16, 24, 20], Among these methods, kernelized and supervised hashing (KSH) [20] is considered very effective, achieving state-of-the-art performance with a moderate training cost. This method is particularly beneficial for bridging the semantic gap in histopathological image analysis. Therefore, this was chosen for scalable image retrieval and classification of breast microscopic images [33]. However, hashing methods tend to generate an unordered set for the same hash value, adversely affecting the classification accuracy when using majority voting. This is particularly true for cell-level analysis of histopathological images, since the differences of cells are very subtle. In the following section, we introduce the weighted hashing to alleviate this problem and it can accurately classify millions of cells. Note that our focus is not to improve the precision of general image retrieval via hashing, but to enhance the classification accuracy for this specific medical image problem.

3. Methodology

3.1. Overview

Fig. 1 shows the overview of our proposed framework, which includes offline learning and online classification. During offline learning, our system automatically detects and segments all cells from thousands of images, resulting in half million of cell images. Regarding cell detection and segmentation, we propose to improve the single-pass voting (SPV) scheme [25]. Our improvement focuses on handling variations in shape and cell size. Then, texture features are extracted from these cell images and are compressed as binary codes, i.e., tens of bits. These compressed features are stored in hash table for constant-time access even among millions of images.

During online classification, our system segments all cells from a testing image, and same types of fea-

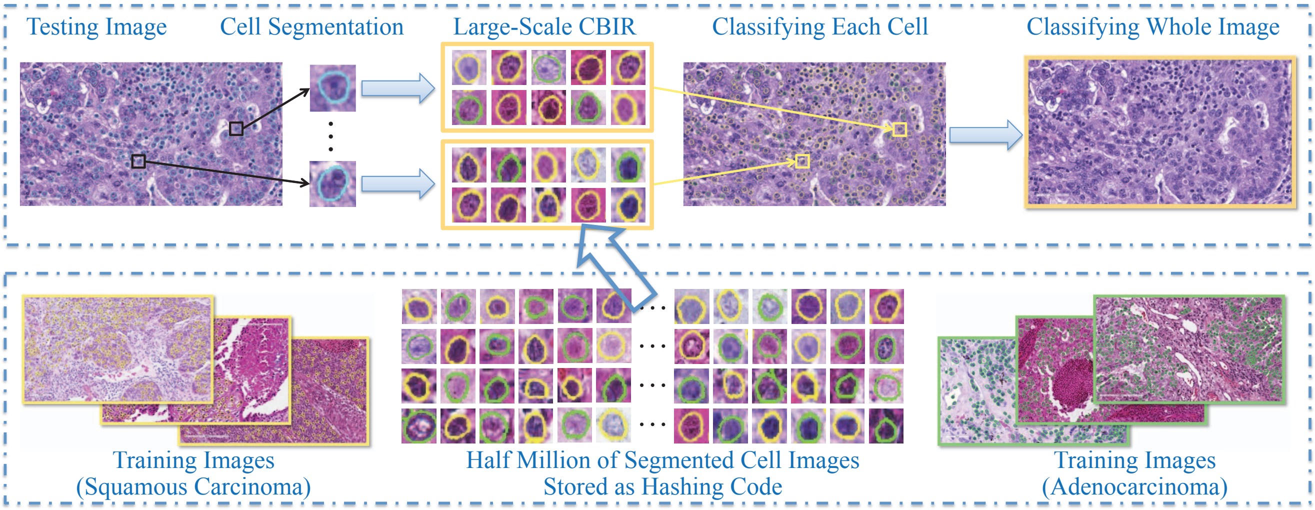

Figure 1. Overview of our proposed framework, based on robust cell segmentation and large-scale cell image retrieval. The top row is the online classification, and the bottom row is the offline learning. Yellow boundaries mean squamous carcinoma, green means adenocarcinoma, and blue means unknown types to be classified.

tures are extracted accordingly and compressed using hashing methods. Then, we perform large-scale cell image retrieval for each segmented cell to classify its category. Finally, the classification result of the testing image is decided by the majority logic, i.e., voting from all cells' classification. Using this scheme, our system can maximally utilize the cell-level information without sacrificing the computational efficiency, owing to the large-scale retrieval via hashing methods. We also design a content-aware weighting scheme to improve the accuracy of traditional hashing methods, based on the observations and priors in histopathological image analysis. In the following sections, we introduce the details of robust cell segmentation and weighted hashing for classification.

3.2. Robust Cell Segmentation

Accurately delineating cells is critical to the cell-level analysis of histopathological images. It includes cell detection and segmentation. The detection algorithm is an improved version of single-pass voting (SPV) proposed by Qi *et al.* [25]. The improvement focuses on handling variations in shape and cell size. The newly introduced 1) region-based hierarchical voting in a distance transform map handles the shape variation, and 2) Gaussian pyramid based voting suppresses the effect of the scale variation. For an image, a Gaussian pyramid is created. At layer l, an SPV is applied with the distance transform being weighted by a Gaussian kernel. Unlike SPV within which each pixel in the voting area receives uniform vote, this weighted voting enables the pixels that locate more inside the cell to receive more votes. Therefore, this mechanism encourages higher voting scores in the central region of the cells. The final vote value is calculated by summing up all the layers:

$$V(x,y) = \sum_{l=0}^{L} \sum_{(m,n)\in S} I[(x,y) \in A_l(m,n)] \cdot C_l(x,y) g(m,n,\mu_x,\mu_y,\sigma), \tag{1}$$

where S denotes the set of all voting pixels, $A_l(m,n)$ denotes the voting area of pixel (m,n) at layer l. $I[\cdot]$ is indicator function, and $C_l(x,y)$ represents the distance transformation map at layer l, $g(m,n,\mu_x,\mu_y,\sigma)$ is an isotropic Gaussian kernel to enhance the robustness of voting. In our experiment, we use Euclidean distance.

The segmentation method is based on active contour [3] with repulsive term. The repulsive term is used to prevent the evolving contours from crossing and merging with each other. Based on the detection result, a circle is associated with each detected cell as initial contour. The ith contour $v_i(s)$ deforms until it achieves a balance between internal force $F^{int}(v_i)$ and external force $F^{ext}(v_i)$ with

$$F^{int}(v_i) + F^{ext}(v_i) = 0, \tag{2}$$

$$F^{int}(v_i) = \alpha v_i''(s) - \beta v_i''''(s), \tag{3}$$

$$F^{ext}(v_i) = \gamma n_i(s) - \lambda \frac{\nabla E_{ext}(v_i(s))}{\|E_{ext}(v_i(s))\|} + \omega \sum_{j=1,j\neq i}^{N} \int d_{ij}^{-2}(s,t) n_j(t) dt, \tag{4}$$

where s indexes the points on the contour, and $v_i''(s)$ and $v_i''''(s)$ with their weights are the second and fourth derivative of $v_i(s)$. $n_i(s)$ with its weight γ denotes the internal pressure force and $\nabla E_{ext}(v_i(s))$ denotes the edges in the image $(\nabla E_{ext}(v_i(s)) = -\nabla\|T[x(s), y(s)]\|^2)$. The last term in 4 represents the repulsive force. N is the number of the neighboring cells. d_{ij} denotes the Euclidean distance between the points of different contours.

This method can robustly detect and segment cells from histopathological images, which are used for cell-level analysis in the next stage.

3.3. Classification via Weighted Hashing

After segmenting all cells from a testing image, our system conducts cell-level classification by exhaustively comparing each cell with all cells in the training database, using hashing-based large-scale image retrieval and majority voting. Hashing has been widely used to compress (high-dimensional) features into binary codes with merely tens of bits [5, 17], allowing mapping into a hash table for constant-time retrieval. However, most traditional hashing methods are not able to provide accurate retrieval or classification in this problem, due to the high intra-class variation of histopathological images. Therefore, supervised information [20] can be leveraged to design discriminative hash functions that are particularly suitable for analyzing histopathological images. The category of each cell is decided straightforwardly with the majority logic of retrieved cells, and the whole image is hence classified by accumulating results of its all cells.

Theoretically, indexing images in a hash table enables constant-time searching, no matter how many samples are used. However, it also requires that the length of binary code is sufficiently short, to store in physical memory for fast access. Given limited number of hash bits, an inevitable limitation is that a large number of images may be mapped into the same hash value. In other words, it may result in an unordered set for the same hash value, where exact or near-exact matches may be obscured within a large-scale database due to noisy features or similar instances. This is particularly true for histopathological image analysis since the differences of cells are very subtle. Consequently, the accuracy of cell classification is adversely affected when choosing the majority of cells mapped into a hash value, and the accuracy of image classification is also limited. Fig. 2 illustrates this inherent limitation of hashing methods in analyzing histopathological images. Half million of cells are mapped into 12 bits, which mean $2^{12} = 4096$ hash values. We visualize the number of cells mapped into each hash value, and the ratio between two types of cells, i.e., adenocarcinoma and squamous carcinoma.

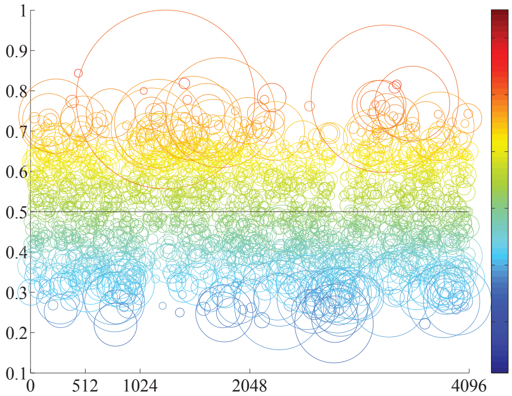

Figure 2. Illustration of the cell distribution in a hash table. X-axis means the hash value using 12 bits, ranging from 0 to 4095, and y-axis means the ratio between two types of cells, ranging from 0 to 1. Each circle means a set of cells mapped to the hash value located in the centroid, its size means the number of cells, and the color map visualizes the ratio of two types of cells, same as the y-axis values.

To solve this inherent problem of hashing methods, we propose a content-aware weighting scheme to reweight the importance of hash values[1] Fig. 2 indicates that cells in certain hash values (i.e., circles in the figure) are not accurate, particularly, circles with around 0.5 ratio, indicating equal chance to be either category. In addition, small sizes of circles are not preferred, since they can be easily affected by many factors, e.g., unusual staining color, inaccurate segmentation results and image noise in our use case. Therefore, we propose to re-weight each hash value by considering its mapped contents. Two metrics are designed, with generalized notations for multi-class classification:

- Support: Given a specific hash value H, the number of cells mapped into H should be considered. This indicates that such amount of cells are used for the classification of this hash value, each with contribution 1, while all remaining cells are irrelevant, i.e., contribution 0. Therefore, we name this metric as “support”, which is conventionally referred to the set of numbers having non-zero values. Denote $S_H = \{\text{cell} : h(\text{cell}) = H\}$ as the set of cells mapping into a specific hash value H, where $h(\text{cell})$ is the hash value of the cell. The support W_H of the hash value H is defined as:

$$W_H = \frac{|S_H|}{\sum_{m=0}^{2^r-1} |S_m|} \tag{5}$$

[1] We use KSH [20] as the base method to generate hash values, because of its efficacy and success in histopathological image analysis [33].

where $|S|$ is the number of element in set S and r is the number of hash bits, representing 2^r hash values.

- Certainty: Instead of assigning a certain category label to each hash value, we should consider the confidence of such categorization and assign a probabilistic label to each hash value. Therefore, this "certainty" term defines the probability of a cell belonging to the ith category when its hash value is H:

$$\begin{aligned} P(L_i|H) &= \frac{P(L_i, H)}{P(H)} \\ &= \frac{|\{\text{cell} : l(\text{cell}) = L_i, \text{cell} \in S_H\}|}{|S_H|} \end{aligned} \quad (6)$$

where $l(\text{cell})$ is the label of a cell image and L_i means the ith label or category.

We combine these two weights to advocate the importance of highly discriminative hash values with sufficient support. Specifically, during the training process, W_H and $P(L_i|H)$ can be computed for all hash values. The category of a whole testing image is decided by:

$$\arg\max_i \sum_{\text{cell}\in\text{query}} W_{H_{\text{cell}}} P(L_i|H_{\text{cell}}) \quad (7)$$

where H_{cell} is the hash value of the cell belonging to the query (testing) image.

This content-aware weighting scheme effectively solves the accuracy issues when using hashing-based retrieval methods for classification. The importance of each cell is decided case-specifically, and accumulating the results of all cells provides accurate classification for the whole image. Regarding the computational efficiency, the overhead during the testing stage lies in the weighted combination, which is negligible as demonstrated in the experiments.

4. Experiments

The lung cancer images were collected from the TCGA dataset [15] and University of Kentucky (Department of Pathology), including 57 adenocarcinoma and 55 squamous carcinoma. 10 patches with 1712 × 952 resolution were cropped from each whole slide scanned pathology specimens. The images were confirmed by three pathologists. 1120 images were used to evaluate the proposed framework. In each image, our algorithm detected and segmented around 430 cells. In total, 484,136 cells were used to evaluate the segmentation accuracy (195,467 adenocarcinoma cells and 288,669 squamous carcinoma cells). All the evaluations were conducted on a 3.40GHz CPU with 4 cores and 16G RAM, in MATLAB and C++ implementation. We validated the efficacy of our proposed framework in terms of both segmentation and classification via image retrieval.

Table 1. Comparative performance evaluation of the detection accuracy. SPV stands for single-pass voting, and PCHT stands for phase-coded Hough transform. MR stands for the missing rate, and OR stands for the over-detection rate.

	Mean	Variance	Min	MR	OR
PCHT	3.7	3.92	0.16	0.46	0.11
SPV	2.9	3.01	0.28	0.21	0.06
Ours	**2.7**	**2.8**	**0.13**	**0.16**	0.08
	FP	TP	Prec	Rec	F_1
PCHT	**0**	0.53	0.995	0.53	0.69
SPV	0.002	0.78	0.996	0.74	0.84
Ours	0.002	**0.83**	**0.997**	**0.84**	**0.90**

4.1. Evaluation of Cell Segmentation

We demonstrate the performance of the cell detection by comparing it with single-pass voting (SPV) and phase-coded Hough transform (PCHT) [32]. We compute the mean, variance and minimum of the deviation of the detected seeds with respect to their ground truth seeds. Note that only the detected seeds within a 8-pixel circle of its ground truth seed are considered. To evaluate the performance more comprehensively, we define a set of metrics including missing rate (MR), over-detection rate (OR), precision (Prec), recall (Rec) and F_1 score. A positive detection is asserted if a detected seed locates within the 8-pixel circle around a ground truth seed, a miss is asserted, otherwise. Over detection is considered as more than one seed are detected in the 12-pixel circle of a ground truth seed. OR is the ratio of the number of such cases over the number of the ground truth seeds. Note that in our experiment, false positive is defined as the case that a seed is detected out of the 8-pixel circle of a ground truth seed yet within its 12-pixel circle. The performance measurements are shown in Table 1.

The performance of the segmentation algorithm is evaluated through comparing our method with four existing methods (mean shift (MS), isoperimetric (ISO) [12], graph-cut and coloring (GCC) [1], and repulsive level set (RLS) [25]), both qualitatively and quan-

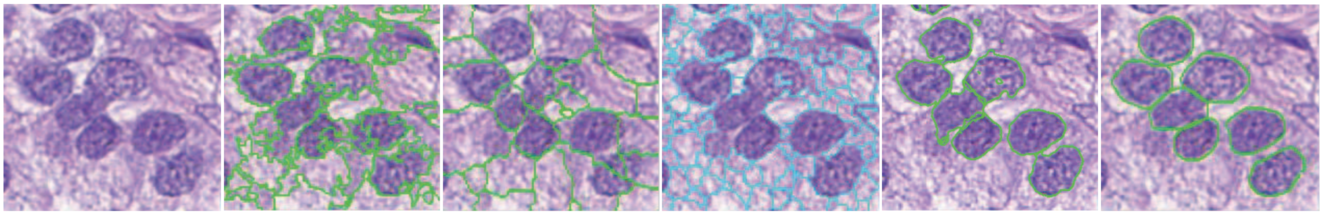

Figure 3. Segmentation results of different methods on a randomly picked patch. From left to right: original image, MS, ISO, GCC, Level Set, and ours.

Table 2. Comparative performance evaluation of the segmentation accuracy for mean shift (MS), ISO [12], GCC [1] and RLS [25]. PM and RM represent precision mean and recall mean. PV and RV denote variances of precision and recall. P80% and R80% denote the sorted highest precision and recall, respectively.

	PM	PV	P80%	RM	RV	R80%
MS	0.73	0.08	0.92	0.79	0.03	0.89
ISO	0.72	0.09	**0.96**	0.81	0.02	0.92
GCC	0.80	0.05	0.95	0.77	0.02	0.89
RLS	0.84	0.02	**0.96**	0.85	0.01	0.92
Ours	**0.87**	**0.01**	0.95	**0.95**	**0.01**	**0.96**

Table 3. Quantitative comparisons of the classification accuracy (the mean value and standard deviation) and running time. Compared methods include kNN [28], DR [26], SVM [6], KSH [20, 33], all of which have been used for histopathological image analysis.

	Adeno	Squam	Mean	STD	Time(s)
kNN	0.309	0.710	0.514	0.049	2605.80
DR	0.458	0.954	0.711	0.082	460.20
SVM	0.929	0.704	0.816	0.080	46.82
KSH	0.861	0.763	0.812	0.079	1.22
Ours	0.887	0.854	**0.873**	0.061	1.68

titatively. The segmentation results of a randomly selected patch are shown in Fig. 3. In our quantitative analysis, we define precision $P = \frac{seg \bigcap gt}{seg}$ and recall $R = \frac{seg \bigcap gt}{gt}$, where seg represents the segmentation result and gt represents the ground truth. We show the mean, variance and 80 % in Table 2.

4.2. Evaluation of Image Classification

In our framework, the image classification (i.e., differentiating adenocarcinoma and squamous carcinoma) is conducted by examining all cells using hashing-based large-scale image retrieval with content-aware weighting. We compare our hashing-based classification scheme with several effective classifiers employed for histopathological image analysis. Following the convention, k-nearest neighbor (kNN) method is used as the baseline to analyze histopathological images [28], because of its simplicity and efficacy. Dimensionality reduction (DR) methods such as principal component analysis (PCA) are effective approaches to improve the computational efficiency and have been employed to analyze histopathological images using high-dimensional features [26]. Support Vector Machine (SVM) is a supervised classification method and widely used in grading systems for breast and prostate cancer diagnosis [6]. We also compare with the KSH [20] that is used as our base method to generate initial hash values. Note that KSH is among state-of-the-art hashing methods and is particularly effective for analyzing histopathological images owing to the non-linearity and supervised information [33]. For fair comparison, same features are used for all compared methods, and their parameters and kernel selections are optimized by cross-validation. Specifically, we use an RBF kernel with optimized gamma value for SVM, and k=9 for kNN.

To conduct the comparison, we randomly select 20% patients as testing data (around 230 images, or 96,000 cells), and use the images from remaining patients as training. This procedure is repeated tens of times to obtain the mean and standard deviation. Table 3 shows the quantitative results of the classification accuracy. Despite the efficacy of kNN in many applications, it fails to produce reasonable results in this challenging problem, due to the large variance of cell images, noise in such large-scale database, and unbalanced number of two classes. DR reduces the feature dimensions, which could be redundant information or noise. The classification accuracy is significantly improved, while still only around 70%. SVM incorporates supervised information, i.e., labels of adenocarcinoma and squamous carcinoma. Not surprisingly, it largely outperforms unsupervised methods, with an accuracy of 81.6%. Although hashing methods are mainly designed for retrieval, i.e., finding approximated nearest neighbors, they are also effective for classification, especially for histopathological images [33]. In our experiment, KSH achieves comparable accuracy as SVM, since it has the same merit of using supervised information. Our proposed hashing method not only utilizes kernels and supervision, but also is equipped with the content-aware weighting scheme to solve the inherent problems of hashing methods. Therefore, it outperforms all other methods, with an accuracy of 87.3%. In addition, the standard deviation of our algorithm is also smaller than other compared methods, indicating the stableness of our algorithm. Table 3 also shows the individual accuracy of adenocarcinoma and squamous carcinoma. Our method achieves the most balanced results for both cases, which is important to this clinical problem as both cases should be recognized and sacrificing the accuracy of one case is not acceptable.

Regarding the computational efficiency of the classification, our hashing method compresses each feature into merely 12 bits, resulting in a hash table with 4096 values, which allow instant access to images mapped into any hash value. Therefore, KSH and our method is real-time, i.e., around 1-2 seconds. Our method uses content-aware weighting and is slightly slower than KSH, due to a small overhead for computing the weighted average. Such computational overhead (i.e., 0.4s) is negligible in practice. Other methods are all significantly slower, ranging from 46 to 2600 seconds. This is the main factor preventing previous methods from being used for cell-level retrieval. Note that the detection and segmentation takes around tens of seconds for each image, and feature extraction takes half second, both of which are the same for all compared methods. The overall speed is quite efficient for practical use.

4.3. Discussions

In this section, we discuss several characteristics of our system for both segmentation and classification. Since the image classification relies on the features extracted from the segmented cells, inaccurate segmentation may adversely affect the classification accuracy. Nonetheless, our system still generates accurate classification results, because of two reasons: 1) Most segmented cells are correct, which is reflected by the high precision and recall. 2) More importantly, the weighting scheme reduces the importance of unreliable features, many of which are extracted from inaccurate segmentations. Particularly, this weighting scheme ensures the robustness of the classification module, making it less sensitive to the segmentation precision. Therefore, our weighted hashing not only benefits the classification accuracy, but also is compatible with the paradigm of cell-level analysis, given the fact that most existing cell segmentation methods are still not perfect.

To demonstrate the generality of our system, we have also conducted preliminary experiments on a dataset of breast cancer images with three classes, including 28 patients for cancer stage I, 37 for II, and 32 for III, with 27,596 segmented cells in total. The accuracy using leave-one-out is 72% for kNN, 75% for DR, 77% for SVM, 78% for KSH and 84% for ours. It is consistent with the lung dataset, indicating the applicability to datasets with multi-class.

This hashing-based classification has one important parameter, i.e., the number of hash bits. Our model is able to generate accurate results within a certain range of parameter values, i.e., not that sensitive to parameters, making it suitable for the large-scale analysis. Furthermore, our content-aware weighting scheme can consistently improve the hashing method for classification accuracy, even with different number of hash bits.

5. Conclusions

In this paper, we proposed a robust and efficient framework to do fine-grained analysis of histopathological images. This is achieved by segmenting all cells and discovering the most relevant instances for each cell among a large database. The main contribution of this proposed framework is to enable scalable and cell-level analysis of histopathological images, as a benefit of our weighted hashing-based classification. This weighting scheme alleviates several intrinsic problems of traditional hashing methods for classification. It significantly improves the diagnosis accuracy of a challenging clinical problem, i.e., differentiating two types of lung cancers as the adenocarcinoma and squamous carcinoma using histopathological images. We envision that it can provide useable tools to assist clinicians' diagnoses of histopathological images and support efficient data management. Although we focused on histopathological image analysis in this paper, the proposed framework is actually very general and potentially applicable to other use cases of medical image analysis, by segmenting region-of-interests such as organs, and retrieving relevant cases from the training database. We plan to investigate this as part of the future work.

This work is supported in part by NIH R01 AR06547901A1 and the funds from Oak Ridge Associated Universities (ORAU) and Charlotte Research Institute (CRI).

References

[1] Y. Al-Kofahi, W. Lassoued, W. Lee, and B. Roysam. Improved automatic detection and segmentation of cell nuclei in histopathology images. *TBME*, 57(4):841 – 852, apr. 2010.

[2] A. N. Basavanhally, S. Ganesan, S. Agner, J. P. Monaco, M. D. Feldman, J. E. Tomaszewski, G. Bhanot, and A. Madabhushi. Computerized image-based detection and grading of lymphocytic infiltration in her2+ breast cancer histopathology. *TBME*, 57(3):642–653, 2010.

[3] L. D. Cohen. On active contour models and balloons. *CVGIP: Image understanding*, 53(2):211–218, 1991.

[4] D. Comaniciu, P. Meer, and D. J. Foran. Image-guided decision support system for pathology. *MVA*, 11(4):213–224, 1999.

[5] M. Datar, N. Immorlica, P. Indyk, and V. S. Mirrokni. Locality-sensitive hashing scheme based on p-stable distributions. In *SoCG*, pages 253–262. ACM, 2004.

[6] S. Doyle, S. Agner, A. Madabhushi, M. Feldman, and J. Tomaszewski. Automated grading of breast cancer histopathology using spectral clustering with textural and architectural image features. In *ISBI*, 2008.

[7] J. G. Dy, C. E. Brodley, A. Kak, L. S. Broderick, and A. M. Aisen. Unsupervised feature selection applied to content-based retrieval of lung images. *TPAMI*, 25(3):373–378, 2003.

[8] S. Edwards, C. Roberts, M. McKean, J. Cockburn, R. Jeffrey, and K. Kerr. Preoperative histological classification of primary lung cancer: accuracy of diagnosis and use of the non-small cell category. *Am J Clin Path*, 53(7):537–540, 2000.

[9] D. J. Foran, L. Yang, et al. Imageminer: a software system for comparative analysis of tissue microarrays using content-based image retrieval, high-performance computing, and grid technology. *JAMIA*, 18(4):403–415, 2011.

[10] D. L. Freeman. Harrison's principles of internal medicine. *JAMA*, 286(8):506, 2001.

[11] Y. Gong, S. Lazebnik, A. Gordo, and F. Perronnin. Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval. *PAMI*, 35(12):2916–2929, 2013.

[12] L. Grady and E. L. Schwartz. Isoperimetric graph partitioning for image segmentation. *T-PAMI*, 28(3):469–475, 2006.

[13] H. Greenspan and A. T. Pinhas. Medical image categorization and retrieval for pacs using the gmm-kl framework. *TITB*, 11(2):190–202, 2007.

[14] M. N. Gurcan, L. E. Boucheron, A. Can, A. Madabhushi, N. M. Rajpoot, and B. Yener. Histopathological image analysis: A review. *TBME*, 2:147–171, 2009.

[15] N. C. Institute. The cancer genome atlas retrieved from https://tcga-data.nci.nih.gov, (2013).

[16] B. Kulis and T. Darrell. Learning to hash with binary reconstructive embeddings. In *NIPS*, pages 1042–1050, 2009.

[17] B. Kulis and K. Grauman. Kernelized locality-sensitive hashing for scalable image search. In *CVPR*, 2009.

[18] G. Langs, H. Müller, B. H. Menze, and A. Hanbury. Visceral: Towards large data in medical imaging - challenges and directions. In *MCBR-CDS MICCAI workshop*, volume 7723 of *Springer LNCS*, 2013.

[19] W. Liu, C. Mu, S. Kumar, and S.-F. Chang. Discrete graph hashing. In *NIPS*, pages 3419–3427, 2014.

[20] W. Liu, J. Wang, R. Ji, Y.-G. Jiang, and S.-F. Chang. Supervised hashing with kernels. In *CVPR*, pages 2074–2081, 2012.

[21] W. Liu, J. Wang, S. Kumar, and S.-F. Chang. Hashing with graphs. In *ICML*, pages 1–8, 2011.

[22] Y. Mu, J. Shen, and S. Yan. Weakly-supervised hashing in kernel space. In *CVPR*, pages 3344–3351. IEEE, 2010.

[23] H. Müller, A. Geissbühler, and P. Ruch. Imageclef 2004: Combining image and multi-lingual search for medical image retrieval. In *MIATSI*, pages 718–727. Springer, 2005.

[24] M. Norouzi and D. M. Blei. Minimal loss hashing for compact binary codes. In *ICML*, pages 353–360, 2011.

[25] X. Qi, F. Xing, D. Foran, and L. Yang. Robust segmentation of overlapping cells in histopathology specimens using parallel seed detection and repulsive level set. *TBME*, 59(3):754 –765, mar. 2012.

[26] O. Sertel, J. Kong, U. V. Catalyurek, G. Lozanski, J. H. Saltz, and M. N. Gurcan. Histopathological image analysis using model-based intermediate representations and color texture: Follicular lymphoma grading. *JSPS*, 55(1-3):169–183, 2009.

[27] Y. Song, W. Cai, and D. Feng. Hierarchical spatial matching for medical image retrieval. In *WMMAR*, pages 1–6. ACM, 2011.

[28] A. Tabesh, M. Teverovskiy, H.-Y. Pang, V. P. Kumar, D. Verbel, A. Kotsianti, and O. Saidi. Multifeature prostate cancer diagnosis and gleason grading of histological images. *TMI*, 26(10):1366–1378, 2007.

[29] J. Wang, S. Kumar, and S.-F. Chang. Semi-supervised hashing for large-scale search. *T-PAMI*, 34(12):2393–2406, 2012.

[30] Y. Weiss, A. Torralba, and R. Fergus. Spectral hashing. In *NIPS*, 2008.

[31] B. Weyn, G. van de Wouwer, A. van Daele, P. Scheunders, D. van Dyck, E. van Marck, and W. Jacob. Automated breast tumor diagnosis and grading based on wavelet chromatin texture description. *Cytometry*, 33(1):32–40, 1998.

[32] Y. Xie and Q. Ji. A new efficient ellipse detection method. In *ICPR*, volume 2, pages 957–960, 2002.

[33] X. Zhang, W. Liu, M. Dundar, S. Badve, and S. Zhang. Towards large-scale histopathological image analysis: Hashing-based image retrieval. *TMI*, 34(2):496–506, 2015.

[34] X. Zhang, L. Yang, W. Liu, H. Su, and S. Zhang. Mining histopathological images via composite hashing and online learning. In *MICCAI*. 2014.