Medical Image Analysis 26 (2015) 306–315

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

High-throughput histopathological image analysis via robust cell segmentation and hashing☆

Xiaofan Zhang[a], Fuyong Xing[b], Hai Su[c], Lin Yang[b,c], Shaoting Zhang[a,*]

[a] Department of Computer Science, University of North Carolina at Charlotte, Charlotte, NC 28223, USA
[b] Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL 32611, USA
[c] Department of Biomedical Engineering, University of Florida, Gainesville, FL 32611, USA

ARTICLE INFO

Article history:
Received 20 October 2014
Revised 13 May 2015
Accepted 16 October 2015
Available online 9 November 2015

Keywords:
Histopathological image analysis
Cell segmentation
Large-scale
Image retrieval
Hashing

ABSTRACT

Computer-aided diagnosis of histopathological images usually requires to examine all cells for accurate diagnosis. Traditional computational methods may have efficiency issues when performing cell-level analysis. In this paper, we propose a robust and scalable solution to enable such analysis in a real-time fashion. Specifically, a robust segmentation method is developed to delineate cells accurately using Gaussian-based hierarchical voting and repulsive balloon model. A large-scale image retrieval approach is also designed to examine and classify each cell of a testing image by comparing it with a massive database, e.g., half-million cells extracted from the training dataset. 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 squamous carcinoma), using thousands of lung microscopic tissue images extracted from hundreds of patients. Our method has achieved promising accuracy and running time by searching among half-million cells .

1. Introduction

Lung cancer is one of the most common cancers in the world (Siegel et al., 2013), and its diagnosis is an extremely important topic for personalized lung cancer treatment. There are four typical histologic types of lung cancers, including adenocarcinoma, squamous carcinoma, small cell carcinoma, and large cell carcinoma, each of which needs a different treatment (Freeman, 2001). Therefore, the accurate staging of lung cancer can help clinicians in determining patient centered treatment, allow for reasonable prognostication, and facilitates comparisons between patient groups in clinical studies. Specifically, current investigations into early detection and adjuvant chemotherapy heavily rely on the proper staging of patients' cancer type. Not only separating small cell carcinoma (SCC) from non-small cell carcinoma (NSCC) is important, it is also strongly recommended (Travis et al., 2011) to subtype NSCC into more specific types such as adenocarcinoma and squamous cell carcinoma, because (1) adenocarcinomas can be tested for epidermal growth factor receptor (EGFR) mutations as a predictor of response to EGFR tyrosine kinase inhibitors; (2) adenocarcinoma response to pemetrexed therapy is better than squamous; (3) potential life-threatening hemorrhage might occur in patients who have squamous cell carcinoma but misclassified and are given bevacizumab. Bronchial biopsy is one of the most effective diagnosis methods to differentiate them, with the aid of Computer Aided Diagnosis (CAD) systems (Kayser et al., 2002; Thunnissen et al., 1992; Mijović et al., 2008). However, most previous methods have emphasized on the diagnosis of small cell vs. non-small cell (i.e., adenocarcinoma, squamous carcinoma, and large cell carcinoma) types of lung cancers. A few efforts have been put on the differentiation of the adenocarcinoma and squamous carcinoma, both of which belong to NSCC, although this task is clinically significant as their management protocols are different (Edwards et al., 2000).

The main challenge of this task is the need of analyzing all individual cells for accurate diagnosis, since the difference between the adenocarcinoma and squamous carcinoma highly depends on the cell-level information, such as its morphology, shape and appearance. In fact, there are a lot of cellular features used by pathologists to differentiate adenocarcinoma from squamous cell carcinoma. Currently, all of them are estimated in a subjective way without rigorous quantifications. These include, but not limited to: (1) nucleoli are often more prominent and obvious in adenocarcinoma tumor cells than squamous cell carcinoma; (2) the individual cell borders tend to be sharper in squamous cell carcinoma than adenocarcinoma; (3) only squamous cell carcinoma contains intercellular bridges; (4) adenocarcinoma has relatively lower nuclear/cytoplasmic ratios and delicate, vacuolated cytoplasm compared with squamous cell carcinoma.

☆ This paper was recommended for publication by Dr. James Duncan.

* Corresponding author. Tel.: +1 7329919820.
E-mail address: rutgers.shaoting@gmail.com, shaoting@cs.rutgers.edu, szhang16@uncc.edu (S. Zhang).

http://dx.doi.org/10.1016/j.media.2015.10.005
1361-8415/

Therefore, rigorously measuring and analyzing each individual cell is important and can assist pathologists for accurate diagnosis. However, a region-of-interest (ROI) image may contain hundreds or thousands of cells. Analyzing each cell can be computationally inefficient using traditional methods. As a result, most previous methods encode the whole image as holistic features by representing the statistics of cell-level information (e.g., architecture features (Doyle et al., 2008) or frequency of local textures (Zhang et al., 2015a)), and may compress high-dimensional features to improve the computational efficiency. Despite the compactness and hence the efficiency, information loss is inevitable in such holistic representation. Therefore, efficiently analyzing each cell is important to investigate. In addition, all the aforementioned cellular features and analysis can only be measured after we complete the accurate cell-level segmentation.

In this paper, we design an automatic framework for the large-scale cell-level analysis of histopathological images, which can examine millions of cells in real-time (preliminary results have been reported in Zhang et al., 2015b). Our solution includes two important modules, robust cell segmentation and large-scale cell retrieval. Specifically, segmentation module provides automatic and robust delineation and measurement of cells, enabling effective feature extraction for each cell. The large-scale image retrieval framework can locate similar instances among massive databases of cells, by improving the efficient hashing methods (Datar et al., 2004; Kulis and Grauman, 2009). Given a new image to be diagnosed, our system automatically segments all cells and efficiently discovers the most relevant cells by comparing them with the training database (e.g., millions of cells extracted from thousands of images). The diagnosis is decided by classifying each cell and using the majority logic. We conduct extensive experiments to differentiate lung cancers, i.e., adenocarcinoma and squamous carcinoma, using a large dataset containing thousands of lung microscopic tissue images acquired from hundreds of patients. Our proposed framework achieves 87.3% accuracy in real-time, by searching a massive database of half million cells extracted from this dataset.

The major contribution of this paper is twofold. (1) A comprehensive and real-time framework is designed to analyze histopathological images by examining all cells. This framework opens a new avenue for investigating large-scale databases, and is particularly suitable for this challenging use case. (2) In terms of technical contribution, 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. Furthermore, we also improve the cell segmentation algorithms by handling variations in shape and cell size, which provide robust and accurate delineations of cells.

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

2. Related work

2.1. Cell segmentation

Various approaches of segmentation in pathological image have been investigated. In Al-Lahham et al. (2012), K-means clustering is used to segment out the cancer cell nuclei at pixel level in a transformed color space. In Loukas et al. (2003), PCA is applied to learn a color space transform and the cell nuclei are segmented out by globally thresholding the transformed image. In Markiewicz et al. (2008), 2009), support vector machine (SVM) classifiers are trained to segment background and the cells based on color or morphological features. Because the above approaches mainly rely on color, they do not work well when there exist non-negligible amount of touching cells present in the images.

Watershed transformation and its variants for splitting touching objects have been widely studied (Vincent and Soille, 1991). A RGB color-based segmentation followed by the watershed algorithm is proposed to tackle the touching cells in Grala et al. (2009), and a 3D watershed algorithm incorporating gradient information and geometric distance of nuclei is represented in Lin et al. (2003). In order to handle over-segmentation, marker-controlled watershed is investigated in Grau et al. (2004), Schmitt and Hasse (2008). In particular, Jung and Kim (2010) developed an H-minima transform based marker-controlled watershed algorithm for clustered nucleus segmentation on histopathological images, and an adaptive H-minima transform is reported in Cheng and Rajapakse (2009) to generate markers for the watershed algorithm. H-minima transform is relatively robust to noise, but it usually requires a careful choice of the h value. Learning based approaches are also exploited to detect markers for watershed algorithms. Mao et al. (2006) applied a supervised marker detection based watershed to cell segmentation on bladder inverted papilloma images, where the markers are located by using a classifier with a combination of photometric and shape information. In Akakin et al. (2012), an SVM classifier is used to automatically detect markers for the watershed algorithm. Compared with unsupervised learning, the supervised marker detection algorithms might provide better performance, but they need sophisticated feature design, which is very challenging due to the complex characteristics of digital pathology images.

Graph-based segmentation methods (Kolmogorov and Zabih, 2004; Boykov and Funka-Lea, 2006) can also be used to automatically segment cells. The nodes of the graph represent pixels or superpixels and each edge corresponds to one pair of neighboring nodes. Image segmentation is achieved by partitioning the graph into several components. Lucchi et al. (2010) exploited a mincut-maxflow algorithm to partition the superpixel based graph, Bernardis and Yu (2010) segmented out individual cells based on the normalized cuts (Shi and Malik, 2000), and (Zhang et al., 2014a) employed a correlation clustering method to achieve superpixel graph partition. Some other graph based methods can be found in Al-Kofahi et al. (2010), Nath et al. (2006), Faustino et al. (2009), Chen et al. (2008), Wu et al. (2012), Yu et al. (2010), Janowczyk et al. (2012) and Lou et al. (2012). Although efficient graph-based segmentation algorithm (Felzenszwalb and Huttenlocher, 2004) is proposed, generally graph partition methods exhibit high time cost, which limits their applications in real cell segmentation.

Deformable models are another popular type of cell segmentation algorithms in biomedical image analysis. A multireference level set algorithm is used for nucleus segmentation in Chang et al. (2012), a dynamic watershed scheme is introduced to the level set model with topology dependence for cell segmentation in Yu et al. (2009), and several repulsive level set approaches are reported in Yan et al. (2008), Ali et al. (2011), Ali and Madabhushi (2012) and Qi et al. (2012). Xu et al. (2007) formulated the active contour model into a graph cut framework, which deforms the contour towards a global minimum within the contour neighborhood. In general, these methods are suitable can naturally handle topology changes, but they might create undesired contours with inhomogeneous regions. Therefore, the parametric active contour models are an alternative approach. Li et al. (2007) applied a gradient flow tracking to 3D nuclei segmentation algorithm, and Cai et al. (2006) developed a repulsive active contour model based on gradient vector flow (GVF) (Xu and Prince, 1998) to segment neuronal axons. However, GVF snake requires clean edge maps to calculate the gradient vector flow, and this might suffer from background clutter in histopathological images.

There exist other types of state-of-the-arts for automatic cell segmentation. Kong et al. (2011) first separated cellular regions from the background with a supervised pixel-wise classification, and then