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.

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. As shown in Fig. 1, 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.

The main technical contribution of this paper is the weighted hashing 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. Specifically, we aim to index millions of cell images in a hash table for constant-time searching, which requires that the length of the 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 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.

To solve this inherent problem of hashing for classification, our solution is a content-aware weighting scheme to re-weight the importance of hash values, generated from KSH [1] (this is chosen as the base method to generate hash values, because of its efficacy and success in histopathological image analysis [2]). 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. Therefore, 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, i.e., the probability of a cell belonging to the ith category when its hash value is H. This can be interpreted as soft assignment on hash values. 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 also re-weight each hash value by considering the number of its mapped cells. 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 testing stage lies in the weighted combination, which is negligible.

The experiment dataset of the lung cancer images was collected from the TCGA dataset 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). Our method outperforms others including KSH, 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. Both 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. To conclude, we proposed a robust and efficient framework to do fine-grained analysis of histopathological images, by segmenting all cells and discovering the most relevant instances for each cell among a large database. We expect that it can provide useable tools to assist clinicians' diagnoses of histopathological images.

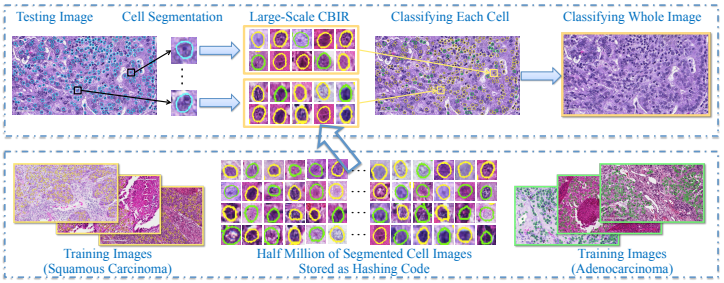

Figure 1: Overview of our proposed framework, based on robust cell segmentation and large-scale cell image retrieval. The top row is the online classification, and the bottom row is the offline learning. Yellow boundaries mean squamous carcinoma, green means adenocarcinoma, and blue means unknown types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

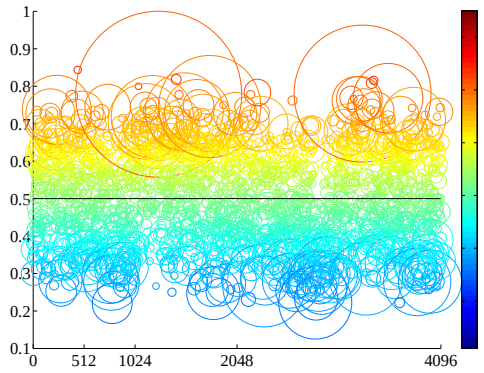

Figure 2: Illustration of the cell distribution in a hash table. X-axis means the hash value using 12 bits, ranging from 0 to 4095, and y-axis means the ratio between two types of cells, ranging from 0 to 1. Each circle means a set of cells mapped to the hash value located in the centroid, its size means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

[1] Wei Liu, Jun Wang, Rongrong Ji, Yu-Gang Jiang, and Shih-Fu Chang. Supervised hashing with kernels. In *CVPR*, pages 2074–2081, 2012.

[2] 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.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.