ABSTRACT

The Cancer Genome Atlas (TCGA) provides a rich repository of whole mount tumor sections that are collected from different laboratories. However, there are a significant amount of technical and biological variations that impede analysis. We have developed a novel approach for nuclear segmentation in histology sections, which addresses the problem of technical and biological variations by incorporating information from manually annotated reference patches with the local color space of the original image. Subsequently, the problem is formulated within a multi-reference graph cut with geodesic constraints. This approach has been validated on manually curated samples and then applied to a dataset of 440 whole mount tissue sections, originating from different laboratories, which are typically 40k-by-40k pixels or larger. Segmentation results, through a zoomable interface, and extracted morphometric data are available at: http://tcga.lbl.gov.

Index Terms— Nuclear segmentation, Nuclear/Background classification, H&E tissue section

1. INTRODUCTION

Tissue histology provides a detailed insight into cellular morphology, organization, and tumor heterogeneity. In tumor sections, it can be used to identify mitotic cells, cellular aneuploidy, and autoimmune responses. More importantly, if tissue morphology and architecture can be quantified on a very large dataset, it will pave the way for constructing databases that are diagnostic, the same way that genome-wide array technologies have identified molecular subtypes and predictive markers. Genome-wide molecular characterization (e.g., transcriptome analysis) has the advantage of standardized tools for data analysis and pathway enrichment, which can enable hypothesis generation in the underlying mechanism. However, the protocol (i) provides an average measurement of the tissue biopsy, (ii) can be expensive, (iii) can hide occurrences of rare events, and (iv) lacks the clarity for translating molecular signature into a phenotypic signature. On the other hand, phenotypic signatures, derived from tissue histology, are hard to compute due to biological and technical variations, but they offer insights into tissue composition and heterogeneity (e.g., mixed populations) and rare events.

In order to have a robust system for characterizing tissue sections, it needs to be able to process samples from multiple laboratories. The Cancer Genome Atlas (TCGA) offers such a collection, where scanned samples originate from different laboratories and are subject to technical (e.g., fixation, staining) and biological (e.g., cell type, cell state) variations. The main technical barrier is that color composition, in the RGB space, is not consistent across tissue sections.

It became clear that a hand segmented dictionary will be needed not only for validation, but also for constructing a model that captures wide variations in the nuclear staining, both within and across tissue sections. Accordingly, our approach integrates local and global image statistics to construct a representation for each pixel based on the Gaussian Mixture Model (GMM). This representation is then regularized with the spatial smoothness constraint through the graph cut framework. The net result is a binarized image of blobs (a single nucleus or a clump of nuclei), which are either validated or partitioned further through geometric reasoning.

Organization of the rest of this paper is as follows: Section 2 reviews previous research; Section 3 describes the details of our approach; Section 4 provides experimental and validation results; and Section 5 concludes the paper.

2. REVIEW OF PREVIOUS WORK

The main issues that hinder correct nuclear segmentation are technical (e.g., sample preparation) and biological heterogeneity (e.g., cell type). Present techniques have focused on adaptive thresholding followed by morphological operators [1], fuzzy clustering [2], level set method using gradient information [3], graph cut method combined with seeds detection[4], color separation followed by optimum thresholding and learning [5], hybrid color and texture analysis that are followed by learning and unsupervised clustering [6]. It is also a common practice that through color decomposition, nuclear regions can be segmented using the same techniques that have been developed for fluorescence microscopy. However, none of these methods can effectively address analytical requirements of the tumor characterization. Thresholding and clustering assume constant chromatin content for the nuclei in the image. In practice, there is a wide variation in chromatin content. In addition, there is the issue with overlapping and clumping of the nuclei, and sometimes, due to the tissue thickness, they cannot be segmented.

One of the main limitations of the above techniques is that they are often applied to a small dataset that originated from a single laboratory. Therefore, some of the inherent variabilities are minimized.

3. APPROACH

Our approach consists of two components: classification between nuclei/background, and nuclear blob partition, as shown in Figure 1. For classification, we leverage both global and local image statistics, in which global image statistics, in both RGB space and LoG...
In this transformed space, the peak of the intensity distribution can be corrected by a simple comparison to the peak of the intensity histogram function, $\text{hist}(R_k) \cdot \text{hist}(N_I_k)/(||\text{hist}(R_k)||||\text{hist}(N_I_k)||)$. Where $\text{hist}(\cdot)$ is the histogram function, $R_k$ is the $k^{th}$ reference image, $N_I_k$ is the normalized input Image $I$ with respect to $R_k$. Then the global fitness term is defined as,

$$E_{gf}(x_p = i) = -\sum_{k=1}^{N} \lambda k^k \log(p_i^k(f^k(p)))$$

(2)

$$= -\alpha \cdot \sum_{k=N+1}^{2N} \lambda k^{k-N} \log(p_i^k(f^k(p)))$$

Where the first and second terms integrate normalized color features and $LoG$ responses, respectively.

### 3.2. Graph Cut Model

Within the graph cut formulation, an image is represented as a graph $G = (V, \tilde{E})$, where $V$ is the set of all nodes, and $\tilde{E}$ is the set of all arcs connecting adjacent nodes. Nodes and edges correspond to pixels ($P$) and their adjacency relationship, respectively. Additionally, there are special nodes that are known as terminals, which correspond to the set of labels that can be assigned to pixels. In the case of a graph with two terminals, terminals are referred to as the source (S) and the sink (T). The labeling problem is to assign a unique label $x_p$ for each node $p \in V$, and the image cutout is performed by minimizing the energy:

$$E = \sum_{p \in V} (E_{data}(x_p) + \gamma E_{smoothness}(x_p, x_q)) + \beta \sum_{(p,q) \in E} E_{smoothness}(x_p, x_q)$$

(1)

where $E_{data}$ is the global data fitness term encoding the fitness cost for assigning $x_p$ to $p$; $E_{smoothness}$ is the local data fitness term encoding the fitness cost for assigning $x_p$ to $p$; $E_{smoothness}$ is the prior energy, denoting the cost when the labels of adjacent nodes, $p$ and $q$, are $x_p$ and $x_q$, respectively; $\beta$ is the weight for $E_{smoothness}$; $\gamma$ is the weight for $E_{data}$. Construction of each of these terms is described as follows:

#### 3.2.1. Global fitness term

The global fitness is established based on manually annotated reference images. Let’s assume $N$ reference images: $R_i, i \in \{1, \ldots, N\}$, and for each reference image, Gaussian Mixture Models are used to represent nuclear and background regions in both RGB space and Laplacian of Gaussian ($LoG$) response space, respectively: $GMM_{Nuclei}, GMM_{Background}$, in which $k \in \{1, \ldots, 2N\}$.

An input test image $I$ is first normalized [11] with respect to every reference image, $R_i$, represented as $N_I_i$. Subsequently, $LoG$ responses of $N_I_i$ are collected to construct $2N$ features per pixels, where the first $N$ features are from the normalized color space, and the last $N$ features are $LoG$ response on the normalized image. Let (i) $f^k(p)$ be $k^{th}$ feature of node $p$; (ii) $\alpha$ be the weight of $LoG$ response; (iii) $p_i^k$ be the probability function of $f^k$ of region $i$ with $i = 0$ : background; $i = 1$ : nuclei; (iv) $\gamma = \sum_{i=0}^{2N} \lambda_i^k \gamma_i^k(p)$; and (v) $\gamma_i^k$ be the weight for $R_i$: $\lambda_k = hist(R_k) \cdot hist(N_I_k)/(||hist(R_k)||||hist(N_I_k)||)$. Where $hist(\cdot)$ is the histogram function, $R_k$ is the $k^{th}$ reference image, $N_I_k$ is the normalized input Image $I$ with respect to $R_k$. Then the global fitness term is defined as,

$$E_{data}(x_p) = -\sum_{k=1}^{N} \lambda^k \log(p_i^k(f^k(p)))$$

(2)
3.2.2. Local Fitness Term

While global fitness term utilizes both color and LoG information in the normalized color space, it does not utilize information in the original color space of the input image. As a result, local variation may be lost, i.e., nuclei having a wide range of chromatin content. The local data fitness is computed as follows:

1) Seeds detection: This step aims to collect local nuclei/background seeds. It incorporates local and global image statistics for improved seed detection. A typical end result is shown in Figure 3(a). The protocol consists of two steps:

- Detect Seeds: Apply the LoG filter (with scale σ) on blue ratio image, detect peaks, and construct a distribution of blue ratio intensity at the peaks corresponding to the negative and positive LoG responses. A small subset of seeds can be mislabeled, where some can be corrected in the following steps.

- Refine seeds: Filtering of seeds (e.g., peaks of the LoG response) are constrained by three criteria: (i) the LoG responses must be above a minimum conservative threshold for background peaks, as shown in Figure 3(b).

II) Local Nuclei/Background color modeling: For each pixel, a local neighborhood is represented by two Gaussian Mixture Models in the original color space. The GMM model is computed from the LoG seeds that are detected in a local neighborhood around p.

The local fitness term is defined as:

\[
E_{l}(x_{p} = i) = -\log(p_{i}(f(p)))
\]

where \( f(p) \) refers to RGB feature of node p in the original color space, and \( p_{i} \) is the probability function of \( f \) of region i (here, \( i = 0 \): background; \( i = 1 \): nuclei), and \( p_{i}(p) = \frac{\sum_{j=0}^{k} GM_{M}(p)}{GM_{M}(p)} \).

3.2.3. Smoothness Term

In order to utilize the gradient information of nuclear boundaries, we adopt the setup from [12], in which the n-links are specifically designed to carry the geodesic information of the input image. Taken a 2D image grid as an example, as shown in Figure 4, the n-link edge weight for \( k^{th} \) family of edge line at node p will be:

\[
w_{k}(p) = \frac{\delta^{2} \cdot |e_{k}|^{2} \cdot \Delta \phi_{k} \cdot \text{det}D(p)}{2 \cdot (e_{k} \cdot D(p) \cdot e_{k})^{T}}
\]

where, \( e_{k} \) is the \( k^{th} \) vector in the neighborhood system, \( \delta \) is the cell-size of the grid, \( \Delta \phi_{k} \) is the angular difference between the \( k^{th} \) and \( (k + 1)^{th} \) edge lines, \( \Delta \phi_{k} = \phi_{k+1} - \phi_{k} \), and \( D(p) \) is a metric continuously varying over points p in a 2D Riemannian space, which is defined as:

\[
D(p) = g(\nabla I) \cdot I + (1 - g(\nabla I)) \cdot u \cdot u^{T}
\]

where \( u = \frac{\nabla I}{|\nabla I|} \) is a unit vector in the direction of image gradient at point p, \( I \) is the identity matrix, and \( g(x) = \exp\left(-\frac{x^{2}}{2\sigma^{2}}\right) \).

Table 1. Edge weights for the graph construction, where \( \mathbb{N} \) is the neighborhood system.

<table>
<thead>
<tr>
<th>Edge</th>
<th>Weight</th>
<th>For</th>
</tr>
</thead>
<tbody>
<tr>
<td>( p \to S )</td>
<td>( E_{gf}(x_{p} = 1) + E_{hf}(x_{p} = 1) )</td>
<td>( p \in P )</td>
</tr>
<tr>
<td>( p \to T )</td>
<td>( E_{gf}(x_{p} = 0) + E_{hf}(x_{p} = 0) )</td>
<td>( p \in P )</td>
</tr>
<tr>
<td>( w_{k}(p, q) )</td>
<td>( w_{k}(p) )</td>
<td>( {p, q} \in \mathbb{N}, \phi_{k} \in {\phi_{k}, \pi + \phi_{k}} )</td>
</tr>
</tbody>
</table>

4. EXPERIMENTAL RESULTS AND DISCUSSION

In order to capture the technical variation, we manually selected and annotated 20 GBM samples (20X), as reference images from TCGA repository. Each sample is a 1k-by-1k block, and an example is shown in Figure 5. For each input image (20X), to be segmented, only top \( M = 10 \) reference images with highest \( \lambda \) were used. The number of components for \( GM_{M} \) was fixed to be 20, and other parameters settings were: \( \alpha = 0.1, \beta = 10.0, \gamma = 0.1, \mu = 10.0, \sigma = 4.0 \) and \( w = 100 \), in which \( \sigma \) was determined based on the preferred dimensions of malignant and normal nuclear size at 20X, and all other parameters were selected to minimize the cross validation error. Two-fold validation was applied on the reference images, and comparisons of average classification performance and segmentation performance were made between our current approach (MRGC) and
Table 2. Comparison of average classification performance between MRGC, and previous approach [13].

<table>
<thead>
<tr>
<th>Approach</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>MRGC</td>
<td>0.79</td>
<td>0.78</td>
</tr>
<tr>
<td>Previous Approach</td>
<td>0.78</td>
<td>0.65</td>
</tr>
</tbody>
</table>

Table 3. Comparison of average segmentation performance between MRGC, and previous approach [13], in which \( \text{precision} = \frac{\# \text{correctly segmented nuclei}}{\# \text{segmented nuclei}} \), and \( \text{recall} = \frac{\# \text{correctly segmented nuclei}}{\# \text{manually segmented nuclei}} \).


5. CONCLUSION AND FUTURE WORK

We have developed a novel approach for segmenting nuclei in H&E tissue sections. Our approach addresses the problem of technical and biological variations by utilizing both global information from the manually annotated reference images, and the local information from the original color space of the target image. The imposed geodesic constrain helps to improve the accuracy of the nuclear boundary. The experimental results demonstrate the effectiveness of our approach. Our future work will focus on improving the nuclear partition algorithm by incorporating nuclear shape model.

6. REFERENCES


