BATCH-INVARIANT NUCLEAR SEGMENTATION IN WHOLE MOUNT HISTOLOGY SECTIONS

Hang Chang[1]*, Leandro A. Loss*[1]*, Paul T. Spellman*[2]*, Alexander Borowsky*[3] *and Bahram Parvin*[1]

[1]Life Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California U.S.A.
[2] Center for Spatial Systems Biomedicine, Oregon Health Sciences University, Portland, Oregon, U.S.A.
[3]Center for Comparative Medicine, University of California, Davis, California, U.S.A.

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 prognostic, 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

This work was supported by NIH grant R01 CA140663 (bp) and U24 CA1437991 (ps) carried out at Lawrence Berkeley National Laboratory under Contract No. DE-AC02-05CH11231.

response space, are extracted from manually selected and annotated reference patches, and local image statistics are established based on foreground and background seeds within a local neighborhood of the image to be segmented. The information above is then condensed and expressed in terms of Gaussian Mixture Models. Having constructed the model, graphcut framework [7] is utilized to classify nuclear and background content. Finally, delineated blobs are subjected to convexity constraints for partitioning clumps of nuclei [8]. In the rest of this section, we will discuss the details of our work.

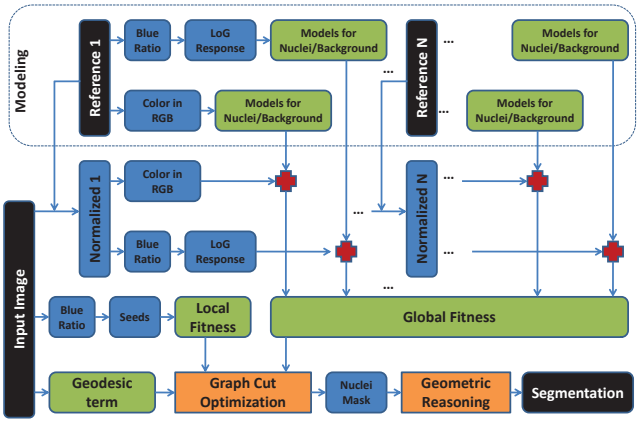


Fig. 1. Steps in Nuclear Segmentation.

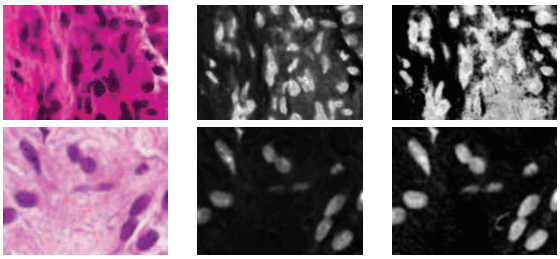
3.1. Color transformation: RGB to Blue Ratio

In order to reduce complexities for integrating LoG responses, the RGB space is transformed to accentuate the nuclear dye. While several techniques for color decomposition have been proposed [10, 9], they are either time-consuming or do not yield favorable outcome as a result of wide technical variations. Our insight led to the following transformation from RGB space into the blue ratio space for computing the LoG responses. $BR = \frac{100*B}{1+R+G} \times \frac{256}{1+B+R+G}$, where B, R and G are the blue, red and green intensities, respectively. Figure 2 demonstrates the immunity of blue ratio to biological and technical variation, compared with the method in [9].

In this transformed space, the peak of the intensity distribution always corresponds to the frequency of occurrence of background pixels. Therefore, some of false negative or positive LoG responses can be corrected by a simple comparison to the peak of the intensity distribution.

3.2. Graph Cut Model

Within the graph cut formulation, an image is represented as a graph $G = \langle \bar{V}, \bar{E} \rangle$, where $\bar{V}$ is the set of all nodes, and $\bar{E}$ is the set of all arcs connecting adjacent nodes. Nodes and edges correspond to pixels ($\mathcal{P}$) 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 $p \in \bar{V}$, and the image cutout is performed by



(a) (b) (c)

Fig. 2. (a) Original images; (b) Blue ratio images; (c) Decomposition by [9].

minimizing the energy:

$$E = \sum_{p \in \bar{V}} \left(E_{gf}(x_p) + \gamma E_{lf}(x_p) \right) + \beta \sum_{(p,q) \in \bar{E}} E_{smoothness}(x_p, x_q) \quad (1)$$

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

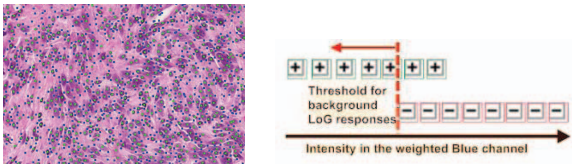
3.2.1. *Global fitness term*

The global fitness is established based on manually annotated reference images. Let's assume N reference images: $R_i, i \in \{1, ...N\}$, 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^k_{Nuclei}, $GMM^k_{Background}$, in which $k \in \{1, ...2N\}$.

An input test image I is first normalized [11] with respect to every reference image, R_i, represented as NI_i. Subsequently, LoG responses of NI_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) α be the weight of LoG response; (iii) $\mathbf{p}^k_i$ be the probability function of f^k of region i with $i = 0 : background; i = 1 : nuclei$; (iv) $\mathbf{p}^k_i(p) = \frac{GMM^k_i(p)}{\sum_{j=0}^{1} GMM^k_j(p)}$; and (v) λ^k be the weight for R_i: $\lambda_k = hist(R_k) \cdot hist(NI_k)/(||hist(R_k)||||hist(NI_k)||)$. Where $hist(\cdot)$ is the histogram function, R_k is the k^{th} reference image, NI_k is the normalized input Image I with respect to R_k. Then the global fitness term is defined as,

$$\begin{aligned} E_{gf}(x_p = i) \quad = \quad & -\sum_{k=1}^{N} \lambda^k log(\mathbf{p}^k_i(f^k(p))) \\ & -\alpha \cdot \sum_{k=N+1}^{2N} \lambda^{k-N} log(\mathbf{p}^k_i(f^k(p))) \end{aligned} \quad (2)$$

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



(a) (b)

Fig. 3. (a) An example of our seeds detection result. Green seeds represent the nuclei, and blue seeds represent the background; (b) LoG responses can be either positive (e.g., potential background) or negative (e.g., foreground or part of foreground) in the transformed blue ratio image. The threshold is set at the minimum intensity in the blue ratio image, which has the most negative LoG response.

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:

I) 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:

1. 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.
2. 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 removing strictly noisy artifacts; (ii) the intensity associated with the peak of the negative LoG responses (e.g., foreground peaks) must concur with the background threshold that is established in Section 3.1; and (iii) within a small neighborhood of $w \times w$, the negative LoG response with the minimum blue ratio, is set as a threshold for background peaks, as shown in Figure 3(b).

II) Local Nuclei/Background color modeling: For each pixel, p, 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:

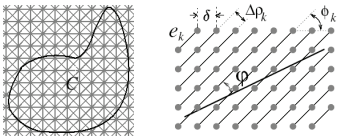
$$E_{lf}(x_p = i) = -log(\mathbf{p}_i(f(p))) \tag{3}$$

where $f(p)$ refers to RGB feature of node p in the original color space, and $\mathbf{p}_i$ is the probability function of f of region i (here, $i = 0 : background; i = 1 : nuclei$), and $\mathbf{p}_i(p) = \frac{GMM_i(p)}{\sum_{j=0}^{1} GMM_j(p)}$

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:

$$w_k(p) = \frac{\delta^2 \cdot |e_k|^2 \cdot \Delta\phi_k \cdot detD(p)}{2 \cdot (e_k^T \cdot D(p) \cdot e_k)^{\frac{3}{2}}} \tag{4}$$



(a) (b)

Fig. 4. (a) Eight-neighborhood 2D grid. (b) One family of lines.

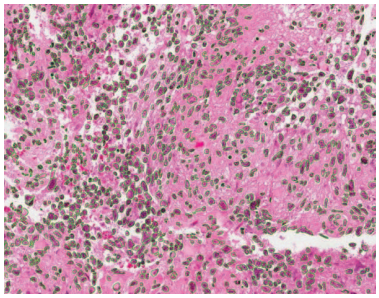


Fig. 5. An example of reference images with manual annotation overlaid as green contours.

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

$$D(p) = g(|\nabla I|) \cdot \mathbf{I} + (1 - g(|\nabla I|)) \cdot \mathbf{u} \cdot \mathbf{u}^T \tag{5}$$

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

Edge	Weight	For
$p \rightarrow S$	$E_{gf}(x_p = 1) + E_{lf}(x_p = 1)$	$p \in \mathcal{P}$
$p \rightarrow T$	$E_{gf}(x_p = 0) + E_{lf}(x_p = 0)$	$p \in \mathcal{P}$
$w_e(p, q)$	$w_k(p)$	$\{p, q\} \in \mathbb{N}$, $\phi_{\overrightarrow{pq}} \in \{\phi_k, \pi + \phi_k\}$

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

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 λ were used. The number of components for GMM was fixed to be 20, and other parameter settings were: $\alpha = 0.1$, $\beta = 10.0$, $\gamma = 0.1$, $\mu = 10.0$, $\sigma = 4.0$ and $w = 100$, in which σ 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

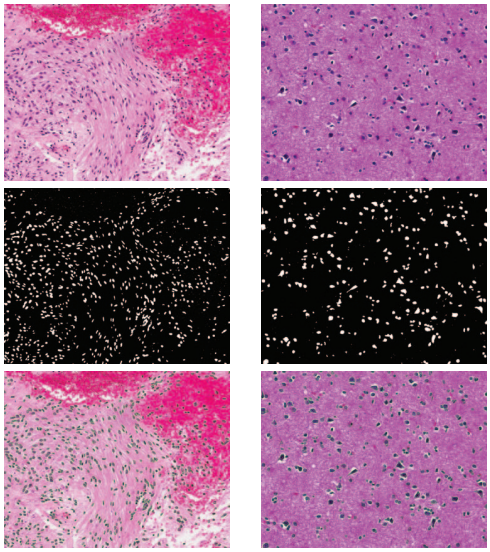


Fig. 6. Row 1: Original images; Row 2: Classification results via MRGC; Row 3: Nuclear partition results via geometric reasoning.

Approach	Precision	Recall
MRGC	0.79	0.78
Previous Approach	0.78	0.65

Table 2. Comparison of average classification performance between MRGC, and previous approach [13].

our previous approach [13], as shown in Table 2 and Table 3, respectively. Having evaluated performance of the system, we applied our method to a large dataset containing 440 GBM tissue sections which are typically 40k-by-40k pixels or larger, and the results were used for integrated analysis [14]. Figure 6 shows some snapshots of the classification and segmentation results; the complete results for all the GBM tissue sections are available at: http://tcga.lbl.gov

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

[1] B. Ballaro, A. Florena, V. Franco, D. Tegolo, C. Tripodo, C. Valenti, An automated image analysis methodology for classifying megakaryocytes in chronic myelprliferative disorders, In Medical Image Analysis 12 (2008) 703–712.

[2] L. Latson, N. Sebek, K. Powell, Automated cell nuclear segmentation in color images of hematoxylin and eosin-stained breast biopsy, In Analytical and Quantitative Cytology and Histology 26 (6) (2003) 321–331.

[3] H. Fatakdawala, J. Xu, A. Basavanhally, G. Bhanot, S. Ganesan, F. Feldman, J. Tomaszewski, A. Madabhushi, Expectation-maximization-driven geodesic active contours with overlap resolution (emagacor): Application to lymphocyte segmentation on breast cancer histopathology, In IEEE Transactions on Biomedical Engineering 57 (7) (2010) 1676–1690.

[4] Y. Al-Kofahi, W. Lassoued, W. Lee, B. Roysam, Improved automatic detection and segmentation of cell nuclei in histopathology images, IEEE Transactions on Biomedical Engineering 57 (4) (2010) 841–852.

[5] H. Chang, R. Defilippis, T. Tlsty, B. Parvin, Graphical methods for quantifying macromolecules through bright field imaging, In Bioinformatics 25 (8) (2009) 1070–1075.

[6] M. Datar, D. Padfield, H. Cline, Color and texture based segmentation of molecular pathology images using hsoms, in: ISBI, 2008, pp. 292–295.

[7] Y.Boykov, M.-P. Jolly, Interactive graph cuts for optimal boundary and region segmentation of objects in N-D images, in: Proc. of IEEE ICCV, 2001, pp. 105–112.

[8] Q. Wen, H. Chang, B. Parvin, A delaunay triangulation approach for segmenting clumps of nuclei, in: ISBI, 2009, pp. 9–12.

[9] A. Ruifork, D. Johnston, Quantification of histochemical staining by color decomposition, Anal Quant Cytol Histology 23 (4) (2001) 291–299.

[10] A. Rabinovich, S. Agarwal, C. Laris, J. H. Price, S. Belongie, Unsupervised color decomposition of histologically stained tissue samples., in: NIPS, 2003.

[11] S. Kothari, J. H. Phan, R. A. Moffitt, T. H. Stokes, S. E. Hassberger, Q. Chaudry, A. N. Young, M. D. Wang, Automatic batch-invariant color segmentation of histological cancer images., in: ISBI, IEEE, 2011, pp. 657–660.

[12] Y. Boykov, V. Kolmogorov, Computing geodesics and minimal surfaces via graph cuts, in: Proc. of IEEE ICCV, 2003.

[13] H. Chang, G. Fontenay, J. Han, G. Cong, F. Baehner, J. Gray, P. Spellman, B. Parvin, Morphometric analysis of tcga gliobastoma multiforme, BMC Bioinformatics 12 (1).

[14] J. Han, H. Chang, G. V. Fontenay, P. Spellman, A. Borowsky, B. Parvin, Molecular bases of morphometric composition in glioblastoma multiforme, in: ISBI, 2012.

Approach	Precision	Recall
MRGC	0.75	0.85
Previous Approach	0.63	0.75

Table 3. Comparison of average segmentation performance between MRGC, and previous approach [13], in which $precision = \frac{\#correctly_segmented_nuclei}{\#segmented_nuclei}$, and $recall = \frac{\#correctly_segmented_nuclei}{\#manually_segmented_nuclei}$.