Quantitative Analysis of Histological Tissue Image based on Cytological Profiles and Spatial Statistics

Young Hwan Chang, Guillaume Thibault, Vahid Azimi,
Brett Johnson, Danielle Jorgens, Jason Link, Adam Margolin, Joe W. Gray

***Abstract*— The cellular heterogeneity and complex tissue architecture of most tumor samples is a major obstacle in image analysis on standard hematoxylin and eosin-stained (H&E) tissue sections. A mixture of cancer and normal cells complicates the interpretation of their cytological profiles. Furthermore, spatial arrangement and architectural organization of cells are generally not reflected in cellular characteristics analysis. To address these challenges, first we describe an automatic nuclei segmentation of H&E tissue sections. In the task of deconvoluting cellular heterogeneity, we adopt Landmark based Spectral Clustering (LSC) to group individual nuclei in such a way that nuclei in the same group are more similar. We next devise spatial statistics for analyzing spatial arrangement and organization, which are not detectable by individual cellular characteristics. Our quantitative, spatial statistics analysis could benefit H&E section analysis by refining and complementing cellular characteristics analysis.**

I. Introduction

In the task of grading or diagnosis of cancer in histopathology images, the identification of certain histological structures such as cancer nuclei, lymphocytes, and glands is essential. For example, cell counts may have diagnostic significance for some cancerous conditions [1], [2]; a low Gleason score means that the cancer tissue is similar to normal prostate tissue and the tumor is less likely to spread. In addition, in [3], the authors found that stromal features are significantly associated with survival and these findings implicate stromal morphologic structure as a previously unrecognized prognostic determinant for breast cancer.

Therefore, the shape, size, extent and other morphological appearance of these structures can be used as indicators for presence or grade of disease and thus, it is important to have the ability to automatically identify these structures. In the past decade, the development of generic and robust cell segmentation methods has intensified [4]. Also, many automated cell image analysis methods have been proposed which allow accurate identification and quantitative measurement of cells' features [5]. Despite these advances, general cellular heterogeneity has remained a significant bottleneck in automated cell image analysis. Recently, many machine learning approaches have been used for automated cell classification by selecting and combining multiple features [3], [5] but they require the segmented cells assessment by a pathologist visually examining individual cells, which is time-consuming and often infeasible for large-scale studies.

The authors are with the Department of Biomedical Engineering, Oregon Health and Science University (OHSU) Portland, OR 97239 USA chanyo@ohsu.edu

Here we describe our approach for quantitative analysis on H&E tissue sections: (a) an automatic unsupervised segmentation, (b) measurement of multiple features for individual nuclei, (c) effectively clustering them based on their measured features and (d) analyzing the spatial arrangement and organization based on spatial statistics which has not previously been considered. Unlike other approaches, our method is fully automatic and requires no label information. We validate the proposed segmentation algorithm by comparing segmentation result to the ground truth immunofluorescence marker (DAPI) and also demonstrate that spatial statistics could complement cellular characteristics analysis by distinguishing different spatial arrangements along the different cell types.

II. Related Literature

In general, the analysis of H&E sections can be divided into mainly two different approaches [1], [2]: some researchers advocate nuclei segmentation and classification; other groups focus on patch level analysis (e.g., small regions) for tumor representation.

A. Local, Structural Segmentation

The problem of cell segmentation has received increasing attention in past years and several automated cell segmentation methods have been proposed [4]. Most methods use a few basic algorithms for cell segmentation such as intensity thresholding, filtering, morphological operations, region accumulation or deformable models [2]. The majority of these approaches treat microscopy images as general natural images. Also, methods proposed in recent times are often merely new combinations of the existing approaches, but these approaches are limited to a specific application.

B. Large Scale (patch-level) Analysis

Some researchers focus on patch level analysis for tumor representation and classification of histology sections. Image patch classification is an important task in many different medical imaging applications. For example, in [6], the authors propose the use of image features for discriminating epithelium and stroma in histological sections. In [7], the authors perform image patch classification to differentiate various lung tissue patterns. These methods are mostly focused on feature design including texture features, object-level features, and graphs features. Also, various classifiers (Bayesian, k-nearest neighbors, support vector machine, etc.) are investigated in a supervised fashion with labeled data.

III. Method

A. Nuclei Segmentation

The H&E staining method colors cells nuclei blue by hematoxylin, and the nuclei staining is followed by counter-staining with eosin, which colors other structures in various shades of red and pink [8]. Thus, each pixel has intensity (R,G,B) and represents a part of morphological features. In order to segment nuclei, we need to extract useful morphological features from the image and then cluster individual pixels based on their features. To do so, we use a set of Gabor filters with different frequencies and orientations [9], which are particularly appropriate for texture representation and discrimination, i.e., edge detection in image processing.

We stack various features such as intensities and Gabor filters' impulse responses where these features can be chosen by users. Once we map each image pixel to a point in an n-dimensional feature space as shown in Figure 1(left), each pixel is enhanced by chosen features and then, by clustering neighboring pixels which have similar features (i.e., k-means clustering), one can differentiate between foreground and background, or between different tissues and cells or nuclei. Thus, nuclei segmentation can be effectively performed by partitioned groups. Finally, we can also exclude unusual segmented parts based on cytological profiles described in the next section.

B. Cytological Profiling and Clustering

Once we segment individual nuclei in the H&E section based on the pixel level, we can extract cytological profile for individual cells. This cytological profile consists of a set of numbers that describe the cellular characteristics including size, nucleus shape, the intensity and texture of various stains, and thus it can be used for classifying cellular types. For example, in Figure 1(right), different nuclei classes show various textural and morphological characteristics. To obtain morphological characteristics, we measure various features including area, major/minor axis length, perimeter, equivalent diameter, shape indices (eccentricity, Euler number, extent, solidity, compactness, circularity, aspect ratio, etc.) and intensity [5]. Combining these features derives high-dimensional feature vector to describe the characteristics of individual nuclei.

Once we measure these features from the individual nuclei, we use a Landmark-based Spectral Clustering (LSC) [10] for large scale clustering. Let $X = [x_1, \cdots, x_N] \in \mathbb{R}^{m \times N}$ be the data matrix where x_i represents a feature vector corresponding to the i-th nucleus, m represents the dimensionality of a feature vector x_i and N is the number of segmented nuclei. By using sparse coding [11], one can find two matrices: a set of basis vectors $U \in \mathbb{R}^{m \times p}$ and the sparse representation with respect to the basis for each data point $Z \in \mathbb{R}^{p \times N}$ whose product can best approximate $X \approx UZ$. However, solving the optimization problem with sparse constraint is very time consuming and the high computational complexity has limited its applicability. For example, the histological images we work with consist of tens of thousand or hundreds of thousands nuclei. On the other hand, LSC [10] selects a few representative data points as the landmarks so one can treat the basis vectors as the landmark points from a data set. Then, we represent the original data points as the linear combinations of these landmarks and the spectral embedding of the data can be efficiently computed with the landmark-based representation. Thus, we can cluster individual segmented nuclei into different types based on their characteristics.

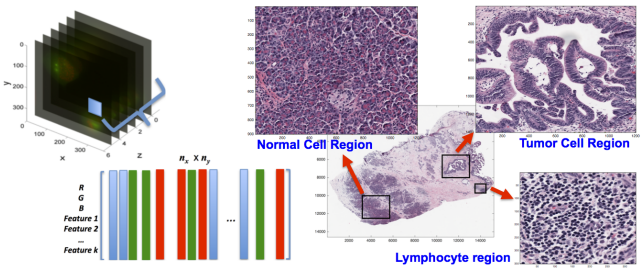

Fig. 1. (left) Conceptual diagram of nuclei segmentation (right) Intratumoral heterogeneity: examples of different classes of cell nuclei (tumor cells / normal cells / lymphocytes).

C. Spatial Statistics [12]

In the previous section, we only use individual cytoprofiles for clustering them into different cellular types. Since spatial arrangement and architectural organization of nuclei is generally not reflected in cellular profiles, this rich information is underused. In addition, biological heterogeneities (e.g., cell type), technical variations (e.g., staining, fixation) and high redundancy in the feature representations can degrade the performance of classifier [13]. To address this issue, we use spatial statistics analysis which complements cellular characteristics analysis and is aimed at characterizing spatial distributions across different cell types such as normal, tumor cells or lymphocytes.

Spatial statistics or spatial analysis is concerned with statistical methods that explicitly consider the spatial arrangement of the data [12]. The observations might be spatially correlated (in two dimensions), which should be accounted for in the analysis. A *spatial point pattern* ($\mathbf{S}$) is a set of point locations in a study region $\mathcal{R}$ and the term *event* can refer to any spatial phenomenon that occurs at a point location. The benchmark model for spatial point patterns is called *complete spatial randomness* (CSR). Under CSR, events are distributed independently and uniformly over the study region as shown in Figure 2 (a).

We look at the behavior of spatial patterns in terms of two properties: first-order properties measure the distribution of events in a study region (spatial density) and second-order properties measure the tendency of events to appear clustered, independently, or regularly-space (interaction between events). We investigate the second-order properties by studying the distances between events in the study region:

1) Nearest neighbor distances - G and F distributions: The G-function measures the distribution of distance from an arbitrary **event** to its nearest neighbors (nearest **event**). The empirical cumulative distribution function for the event-

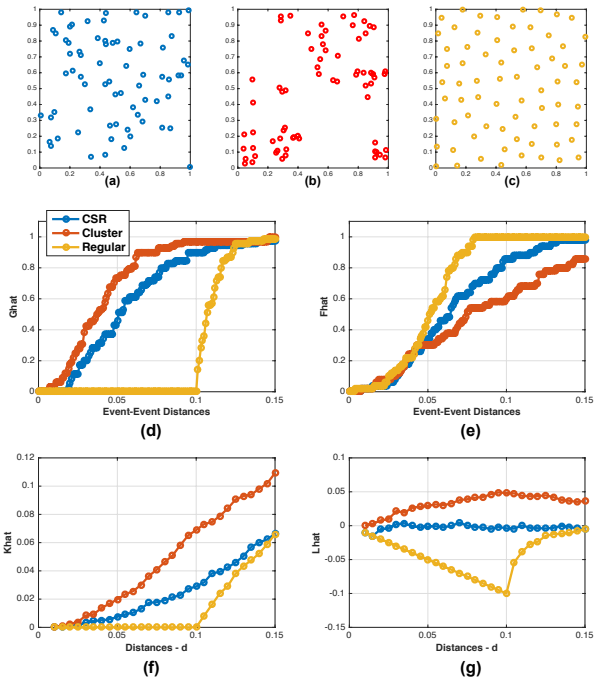

Fig. 2. Examples of spatial point patterns and comparability of a point process with CSR: (a) CSR point process (b) cluster point pattern (c) point pattern exhibiting regularity. Under CSR, an event has the same probability of occurring at any location in $\mathcal{R}$, and events neither inhibit (i.e., regularity) nor attract each other (i.e., clustering) (d-g) G, F, K and L- distributions.

event distances w measures the distribution of distances from an arbitrary event to its nearest neighbors:

$$\hat{G}(w) = \frac{\sum_{i=1}^{n} I_i}{n}, \text{ where } I_i = \begin{cases} 1 \text{ if } d_i \in \{d_i : d_i \leq w, \forall i\} \\ 0 \text{ otherwise} \end{cases}$$

where $d_i = \min_j\{d_{ij}, \forall j \neq i \in \mathbf{S}\}, i = 1, ..., n$. Under CSR, the value of the G-function becomes $G(w) = 1 - e^{\lambda\pi w^2}$ where λ is the mean number of events per unit (intensity). The comparability of a point process with CSR can be assessed by plotting the empirical function $\hat{G}(w)$ against the theoretical expectation $G(w)$ as shown in Figure 2 (d). For instance, for a clustered pattern, observed locations should be closer to each other than expected CSR and thus we would expect that $\hat{G}(w)$ would climb steeply for smaller values of w and flatten out as the distances get larger.

The F-function measures the distribution of all distance from an arbitrary **point** k in the plane to the nearest observed **event** j:

$$\hat{F}(x) = \frac{\sum_{k=1}^{m} I_k}{m}, \text{ where } I_k = \begin{cases} 1 \text{ if } d_k \in \{d_k : d_k \leq x, \forall k\} \\ 0 \text{ otherwise} \end{cases}$$

where $d_k = \min_j\{d_{kj}, \forall j \in \mathbf{S}\}, k = 1, ..., m, j = 1, ..., n$. Under CSR, the expected value is also $F(x) = 1 - e^{\lambda\pi x^2}$. When we examine a plot of $\hat{F}(x)$ (Figure 2 (e)), the opposite interpretation holds. For example, for a clustered pattern, observed locations j should be farther away from random points k than expected under CSR.

2) K, L distributions: A homogeneous set of points in a study region $\mathcal{R}$ is distributed such that approximately the same number of points occurs in any circular region of a given area. A set of points that lacks homogeneity is spatially clustered. A simple probability model for spatially homogeneous points is the Poisson process in $\mathcal{R}$ with constant intensity function. Then, the K-function is defined as $\hat{K}(d) = \lambda^{-1}E[\#\text{extra events within distance } d \text{ of an arbitrary event}]$ where λ is a constant representing the intensity over the region and $E[\cdot]$ denotes the expected value. For a CSR spatial point process, the theoretical K-function is $K(d) = \pi d^2$. Figure 2 (f) shows the function $\hat{K}(d)$ for the data. Note that it is above the curve for a random process (e.g., $\hat{K}(d) > \pi d^2$) indicating possible clustering. Alternatively, if our observed process exhibits regularity for a given value of d, then we expect that the estimated K-function will be less than πd^2.

Another approach, based on the K-function, is to transform $\hat{K}(d)$ using $\hat{L}(d) = \sqrt{\frac{\hat{K}(d)}{\pi}} - d$. Peaks of positive values in a plot of $\hat{L}(d)$ would correspond to clustering and negative values indicating regularity, for the corresponding scale d. In the plot of $\hat{L}(d)$ (Figure 2 (g)), we see possible evidence of clustering at all scales.

IV. EXPERIMENTS AND RESULTS

A. Nuclei Segmentation

In order to quantitatively evaluate the segmentation provided by the proposed method, we compare the segmentation result to the ground truth immunofluorescence marker (DAPI). Figure 3 reports the validation of segmentation result. We also calculate true positive rate (sensitivity) = 0.8070, true negative rate (specificity) = 0.9437, accuracy rate = 0.9249 among 7924 nuclei where we calculate them based on pixel level. Also, the Dice coefficient[1] is 0.7474.

[1]A measure of overlap between two regions, commonly used for evaluation of segmentation techniques, $D(X,Y) = 2\frac{|X \cap Y|}{|X|+|Y|}$.

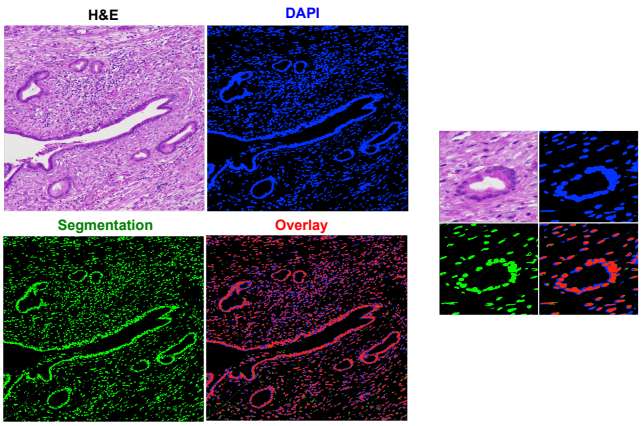

Fig. 3. Validation of segmentation result with matched immunofluorescence: H&E stained section, DAPI (ground truth), segmented nuclei (Segmentation) and overlapped region (Overlay, red color: perfect match, green: only H&E, blue: only DAPI). Note that we only show small region due to the space limit.

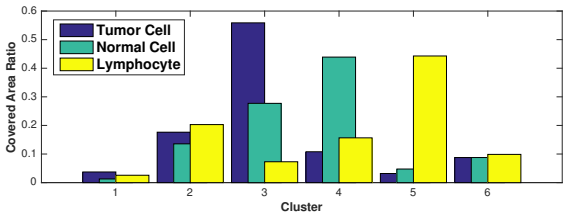

Fig. 4. A covered area ratio subjected to a particular cluster in each H&E stained section. Within the same cluster, cytoprofiles of segmented nuclei show similar characteristics.

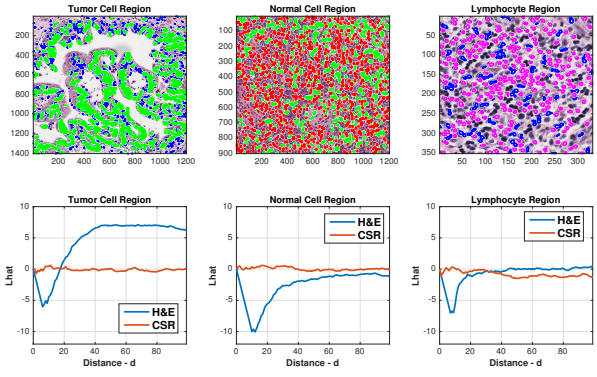

Fig. 5. The segmented nuclei (color-coded according to their clusters) and the second-order spatial statistics (L-function): (left) tumor cell region (cluster 2,3) (middle) normal cell region (cluster 3,4) (right) lymphocyte region (cluster 2,5). Not surprisingly, higher nuclei clustering was found in tumor region compared with normal cell or lymphocyte region, possibly due to the aggregated patterns of tumor cells. Note that we only show two dominant clusters for each region so there are some nuclei which are not color-coded.

B. Quantitative Analysis based on Cellular Characteristics and Spatial Statistics

Once we segment individual nuclei, we extract cellular characteristics from the tumor cell / normal cell / lymphocyte regions as shown in Figure 1(right). In order to characterize different classes of nuclei (among 5431 nuclei), we choose 6 clusters and run LSC. Figure 4 shows a population (covered area ratio) of segmented nuclei subjected to a particular cluster in each H&E section respectively. For example, we observe that nuclei corresponding to cluster 5 and cluster 4 are distinctively dominant in lymphocyte region and normal cell region respectively. However, one cannot perfectly discriminate different classes of nuclei based on cellular characteristics alone. For instance, although nuclei corresponding to cluster 3 are dominant in tumor region, they also exists in normal cell region. Thus, there is no unique cluster representing a specific cell type (i.e., tumor) here.

In order to complement cellular characteristics analysis, we characterize a spatial distribution of dominant nuclei type along the different regions (tumor / normal cell / lymphocyte) and observe that tumor cells are differentially distributed. Figure 5 (top row) shows distribution of individual segmented nuclei which we have color-coded according to their clusters (blue: cluster 2, green: cluster 3, red: cluster 4, magenta: cluster 5) and Figure 5 (bottom row) shows the second-order spatial statistics of selected nuclei. Here, we look at the pattern at several scales, i.e., using $\hat{L}$-function since in general, both $\hat{G}(w)$ and $\hat{F}(x)$ consider the spatial point pattern over the smallest scale which can be a major drawback, especially with clustered patterns where nearest-neighbor distances are very short relative to other distance in the pattern. For a tumor region, we choose dominant types (e.g., cluster 2 and 3) and calculate $\hat{L}$-distribution. As we see a cluster behavior visually, we see strong evidence of clustering in the plot of $\hat{L}(d)$. On the other hand, for both normal cell region and lymphocyte region, we do not see any point pattern exhibits clustering behavior.

V. Conclusions

We have described a simple, but effective methodology for quantitative analysis for H&E section. We demonstrate the performance of the segmentation algorithm by comparing the result to ground truth data (DAPI). Also, we demonstrate that spatial statistics analysis could benefit H&E section analysis by complementing cellular characteristics analysis.

References

[1] M. Gurcan, L. Boucheron, A. Can, A. Madabhushi, N. Rajpoot, and B. Yener, "Histopathological image analysis: A review," *Biomedical Engineering, IEEE Reviews in*, vol. 2, pp. 147–171, 2009.

[2] H. Irshad, A. Veillard, L. Roux, and D. Racoceanu, "Methods for nuclei detection, segmentation, and classification in digital histopathology: A review - current status and future potential," *Biomedical Engineering, IEEE Reviews in*, vol. 7, pp. 97–114, 2014.

[3] A. H. Beck, A. R. Sangoi, S. Leung, R. J. Marinelli, T. O. Nielsen, M. J. van de Vijver, R. B. West, M. van de Rijn, and D. Koller, "Systematic analysis of breast cancer morphology uncovers stromal features associated with survival," *Science Translational Medicine*, vol. 3, no. 108, pp. 108ra113–108ra113, 2011.

[4] E. Meijering, "Cell segmentation: 50 years down the road [life sciences]," *Signal Processing Magazine, IEEE*, vol. 29, pp. 140–145, Sept 2012.

[5] T. R. Jones, A. E. Carpenter, M. R. Lamprecht, J. Moffat, S. J. Silver, J. K. Grenier, A. B. Castoreno, U. S. Eggert, D. E. Root, P. Golland, and D. M. Sabatini, "Scoring diverse cellular morphologies in image-based screens with iterative feedback and machine learning," *Proceedings of the National Academy of Sciences*, vol. 106, no. 6, pp. 1826–1831, 2009.

[6] F. Bianconi, A. Álvarez Larrán, and A. Fernández, "Discrimination between tumour epithelium and stroma via perception-based features," *Neurocomput.*, vol. 154, pp. 119–126, Apr. 2015.

[7] Q. Li, W. Cai, and D. Feng, "Lung image patch classification with automatic feature learning," in *Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE*, pp. 6079–6082, July 2013.

[8] C.-W. Wang, "Robust automated tumour segmentation on histological and immunohistochemical tissue images," *PLoS ONE*, vol. 6, no. 2, 2011.

[9] R. Mehrotra, K. R. Namuduri, and N. Ranganathan, "Gabor filter-based edge detection," *Pattern Recognition*, vol. 25, pp. 1479–1494, December 1992.

[10] D. Cai and X. Chen, "Large scale spectral clustering via landmark-based sparse representation," *Cybernetics, IEEE Transactions on*, vol. 45, pp. 1669–1680, Aug 2015.

[11] N. Nayak, H. Chang, A. Borowsky, P. Spellman, and B. Parvin, "Classification of tumor histopathology via sparse feature learning," in *Biomedical Imaging (ISBI), 2013 IEEE 10th International Symposium on*, pp. 410–413, April 2013.

[12] W. L. Martinez and A. R. Martinez, *Computational Statistics Handbook with MATLAB, Second Edition*. Chapman and Hall/CRC, 2 ed., 2007.

[13] Y. Zhou, H. Chang, K. Barner, P. Spellman, and B. Parvin, "Classification of histology sections via multispectral convolutional sparse coding," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pp. 3081–3088, June 2014.