An Image Analysis Approach for Detecting Malignant Cells in Digitized H&E-stained Histology Images of Follicular Lymphoma

Olcay Sertel[*†], Umit V. Catalyurek[*†], Gerard Lozanski[‡], Arwa Shanaah[‡], Metin N. Gurcan[†]
[*]Dept. of Electrical and Computer Engineering, [†]Dept. of Biomedical Informatics, [‡]Dept. of Pathology, The Ohio State University, Columbus, OH 43210 USA
{osertel, umit}@bmi.osu.edu, {Gerard.Lozanski, Arwa.Shanaah, Metin.Gurcan}@osumc.edu

Abstract

The gold standard in follicular lymphoma (FL) diagnosis and prognosis is histopathological examination of tumor tissue samples. However, the qualitative manual evaluation is tedious and subject to considerable inter- and intra-reader variations. In this study, we propose an image analysis system for quantitative evaluation of digitized FL tissue slides. The developed system uses a robust feature space analysis method, namely the mean shift algorithm followed by a hierarchical grouping to segment a given tissue image into basic cytological components. We then apply further morphological operations to achieve the segmentation of individual cells. Finally, we generate a likelihood measure to detect candidate cancer cells using a set of clinically driven features. The proposed approach has been evaluated on a dataset consisting of 100 region of interest (ROI) images and achieves a promising 89% average accuracy in detecting target malignant cells.

1. Introduction

Recent developments in high-throughput whole-slide digital scanners have accelerated the research on computer-aided diagnosis and prognosis in histopathology. This has the potential to improve the precision and accuracy of the qualitative visual inspection, and assist pathologists in their decision-making mechanism; hence improve the clinical outcomes. In this study, we propose an image analysis approach for histopathological evaluation of follicular lymphoma (FL) tissue samples. FL is a tumor of lymph system and the second most common lymphoid malignancy in the western world [1]. Currently, in clinical and laboratory practice, the gold standard in FL risk stratification is the examination of hematoxylin and eosin (H&E) stained tissue section(s) and performing histological grading [1]. Histological grading is based on the average count of centroblasts (CBs), large malignant cells, in ten random standard high power fields representing follicular regions. However, this is a highly subjective process and the results show well-documented inter- and intra-observer variability for the various grades of FL even among expert pathologists [2, 3].

The proposed image analysis system segments and classifies individual cells in digitized microscopic images of H&E-stained FL tissues samples. For the initial segmentation, we use a robust feature space analysis algorithm, namely the mean shift algorithm to estimate the clusters associated with the modes of the underlying density distribution spanned by the color vector of each pixel. The resulting clusters identified by the mean shift procedure are further pruned to obtain the distributions associated with each sub-cellular component. Further morphological operations are applied to deal with complex sub-cellular composition of individual cells, enforce morphological and spatial constrains and split touching cells. Finally, we construct a feature vector for each segmented cell and detect candidate CBs using the morphology and texture based likelihood functions.

2. Segmentation of individual cells

In order to segment individual cells, we first partition the tissue into basic cytological components, e.g., cell nuclei and cytoplasm, extra cellular material (ECM), red blood cells (RBC) and background. In H&E-stained tissue samples, nuclear and cytoplasm regions have blue-purple color and darker intensity while colors associated with ECM and RBCs are

1051-4651/10 $26.00 © 2010 IEEE
DOI 10.1109/ICPR.2010.76

IEEE computer society

usually in the hues of pink and red. Conventionally, feature space clustering algorithms such as expectation maximization and k-means are widely used for segmentation in histopathology applications [4, 5]. However, these algorithms rely on the prior knowledge of the number of clusters and the implicit assumptions on the shape (mostly elliptical) of these clusters. In fact, especially in histopathology imagery, these assumptions may not comply well with the data due to considerable staining variations between tissue samples; hence the distribution associated with each sub-cellular component should be discerned solely from the image being processed.

In order to achieve an adaptive and robust feature space analysis, we use a non-parametric method, namely the mean-shift algorithm, a procedure for estimating the modes (i.e., stationary points of the density) of a density function given the data sampled from that function [6, 7]. The mean shift algorithm does not require prior knowledge on the number of clusters, and does not constrain the shape of the clusters. Briefly, the mean-shift procedure operates as follows: Let x_i be a set of n points in the d-dimensional Euclidean space R^d that defines the feature space where $i=1,\ldots,n$. Also let $S_h(x)$ define a hyper-sphere of radius h centered on x, containing n_x data points. Then, it can be shown [6] that the sample mean shift vector is:

$$M_h(x) = \frac{1}{n_x}\sum_{x_i \in S_h(x)} x_i - x \cong \frac{\hat{\nabla} f(x)}{\hat{f}(x)}, \qquad (1)$$

where $\hat{f}(x)$ is the estimate of the local density and $\hat{\nabla} f(x)$ is the estimate of the local density gradient. In other words, Eq. 1 states that the mean shift vector always points towards the direction of the maximum increase in the density allowing the local estimate of the normalized gradient to be computed using the sample mean. We recursively compute the mean shift vector and update the hyper-sphere by the mean shift vector until it converges to a local density maximum, which is also referred to as the *mode* of the density. In order to identify all the modes, this procedure is repeated for all the points. Assigning each data point to the mode to which the local density is converged, the underlying clusters with arbitrary shapes in the feature space are identified.

In our application, we construct the feature space by transforming the image from the RGB into the L*u*v* color space, which defines a perceptually uniform color space; therefore we can use the Euclidean distance. Then, we identify the modes of this feature space by applying the mean shift procedure as described above. The only parameter in the mean shift mode estimation process is the radius of the hyper-sphere. We set its value as *h=3.5*, which is determined experimentally by visually observing the results for a range of values (i.e., [2-10]). In our application of segmenting H&E-stained tissue images, after applying the means shift procedure, the resulting number of clusters ranges from 50 to 100, depending on the variations within the image. However, most of these clusters have only a few data samples associated with them, which correspond to insignificant regions mostly due to the imaging artifacts. Therefore, we first eliminated any cluster if the number of data points associated with it corresponds to less than 2% of the total number of data points. The remaining clusters are hierarchically grouped based on similarities in the L*u*v* color space and the spatial transitions between samples in each cluster. The change in intensity and color between nuclei, cytoplasm and ECM can be quite gradual; hence precise segmentation may not be achieved using only the color information. While the color similarity tends to group clusters with similar color vectors, spatial transition probabilities favor to group clusters that are frequently neighboring each other in the spatial domain. Based on the fact that the nuclei component is usually surrounded by the cytoplasm component, and the cytoplasm component is usually surrounded by the ECM component in the spatial domain, the spatial transition probabilities between clusters ensure additional spatial constraints for a more precise segmentation. Accordingly, the pair-wise similarities between the remaining clusters are constructed as follows:

$$S_{ij} = \left|c_i - c_j\right| \cdot \left({}^{1}\!\big/_{t_{ij}} \right) \qquad (2)$$

where c_i, c_j are the i^{th} and j^{th} clusters, and t_{ij} is the spatial transition probability between i^{th} and j^{th} clusters. The t_{ij} measures the frequency of the samples of i^{th} and j^{th} clusters neighboring each other in the spatial domain and it is calculated as follows:

$$t_{ij} = \left|\forall(x,y) \mid I(x,y) = i, I_{N(x,y)}(r,c) = j\right| \qquad (3)$$

where i and j are cluster indices and (x, y) are the image coordinates. The $I_{N(x,y)}$ corresponds to the immediate 8-neighborhood of the pixel denoted by (x, y), and |.| denotes cardinality. Figure 1(a) and (b) show a sample image region and the corresponding color coded segmentation labels after hierarchical grouping, where cell nuclei, cytoplasm, ECM and background are represented in blue, cyan, yellow and white colors, respectively. Figure 1(c) shows the pair-wise similarities between the remaining clusters after the initial pruning step. Figure 1(d) shows the result of hierarchical grouping as a dendrogram plot, where the leaf nodes shown in same colors indicate the grouping.

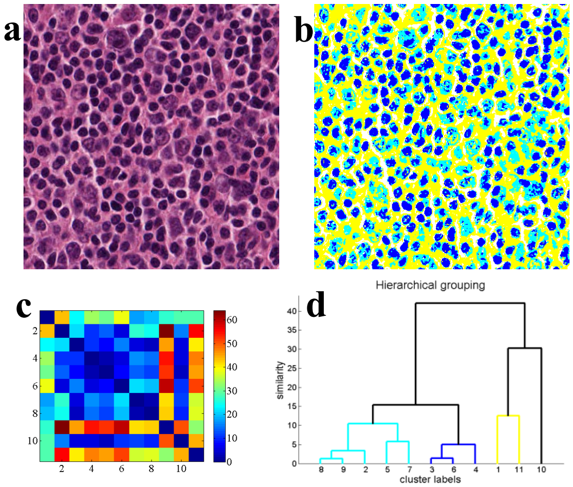

Figure 1. (a) A sample image region and (b) its segmentation result; (c) and (d) show pair-wise similarities and the dendrodram plot of hierarchical grouping.

In order to identify individual cells after the initial segmentation, we used both nuclei and cytoplasm components. The initial cell likelihood map, φ_{init}, is generated by assigning weights to the nuclear and cytoplasm components and applying a Gaussian smoothing that ensures spatial coherence. This is followed by a morphological reconstruction using a top-hat filter to enhance the likelihood image as follows:

$$\varphi_{init} = \left(0.7 * I_{nuc} + 0.3 * I_{cyt}\right) \otimes G_{\sigma=2} \qquad (4)$$
$$\varphi_{enh} = \varphi_{init} - \rho_D(\varphi_{init})$$

where I_{nuc} and I_{cyt} denote the nuclei and cytoplasm components, ⊗ denotes the convolution operation, $G_{\sigma=2}$ is a 7×7 Gaussian mask with a standard deviation $\sigma=2$, and ρ_S indicates a morphological top-hat filter with a disk structuring element. Finally, we applied a marker controlled watershed transformation using the enhanced cell likelihood image, φ_{enh}, to identify the borders between individual cells. Figure 2 shows the results of the post-processing step. As can be seen in Figure 3(d), the segmentation is successful for the vast majority of cells.

3. Classifying CB cells

There are several types of cells in FL tissue (e.g., centrocytes, CBs, macrophages, etc.). The majority of them are centrocytes characterized by compact cells with coarse chromatin and scant cytoplasm. CBs, on the other hand, are characterized by larger size, vesicular chromatin, and multiple prominent nucleoli that are frequently associated with nuclear membrane.

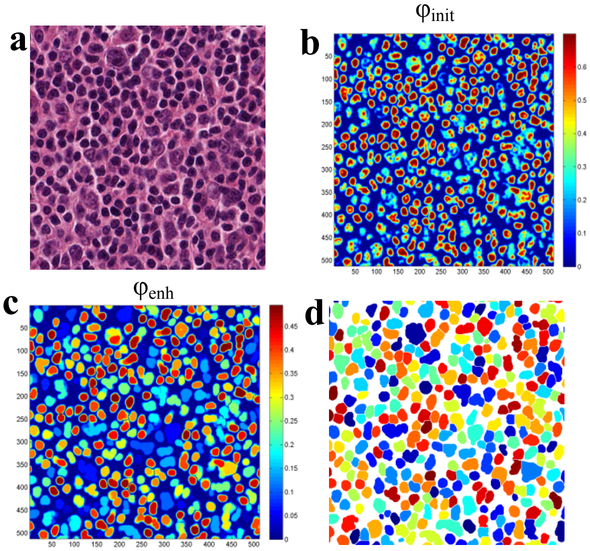

Figure 2. (a) Sample image region, (b-c) steps of the cell segmentation, and (d) the final result, where each cell is labeled in different colors.

Typically, there exist relatively fewer number of CB cells as opposed to centrocytes; hence traditional classification approaches do not yield good results due to underrepresentation of CB samples. Therefore, using a set of features based on prior clinical knowledge, we first constructed a CB likelihood measure. Then, for each image, we adaptively computed a threshold value from the distribution of CB likelihood and generated the candidate CB cells.

The feature vector constructed for each segmented cell includes area, nuclear to cytoplasm ratio, extent of nuclear regions (i.e., the proportion of the pixels in the bounding box that are also in the region) and, mean and standard deviation of the intensity range within the cell. Basically, these features capture morphological characteristics such as size, sub-cellular composition and nuclear shape irregularity, as well as textural characteristics of cells for further detection of CBs. Using prior information based on clinical and biological knowledge and a small set of training samples, we constructed CB likelihood functions for each feature. Figure 3 shows the constructed likelihood functions for the area, nuclear to cytoplasm ratio and extent features. Assuming independence between these features, the final CB likelihood, Λ_{CB}, is calculated as follows:

$$\Lambda_{CB} = \prod_{\forall \lambda_i} \lambda_i \qquad (5)$$

where λ_i is the CB likelihood of the i^{th} feature measure.

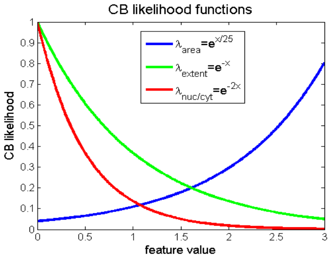

Figure 3. CB likelihood functions of the area, nuclear to cytoplasm ratio and extent

4. Experimental results and evaluation

For the computerized CB detection, we applied an adaptive threshold to the generated CB likelihood Λ_{CB}. The threshold value, τ, is computed as follows:

$$\tau = \arg\max_{x} \quad CMF_{\Lambda_{CB}}(x) < 0.95 \qquad (6)$$

where $CMF_{\Lambda_{CB}}$ is the cumulative mass function of Λ_{CB}. The value 0.95 is determined empirically based on the fact that there is limited number of CB cells as opposed to the large number of non-CB cells in the FL tissue samples. Figure 4(a) shows the CB likelihood computed for the sample image region shown in Figure 2. Figure 4(b) shows the CB detection results after applying the adaptive thresholding. As can be seen in this sample ROI image, all of the CBs are detected with a few false positive detections.

We validated and evaluated the proposed approach over a dataset of 100 region of interest (ROI) images captured from ten whole-slide tissue samples digitized at 40× magnification using an Aperio Scope XT scanner (Aperio, San Diego, CA, USA). The resolution of each ROI image is 1353×2168 pixels, approximately including 2200 cells. The ground-truth information consists of the locations of individual CB cells in each ROI image generated by two expert board-certified hematopathologists. The accuracy of the proposed system is computed by comparing the centroid locations of detected CB cells with the ground-truth CB locations. Table 1 shows the evaluation results, where we report the average percentage of correctly identified CBs and the false positive detection rate as well as the percentage of correctly identified non-CB cells. Due to the agreement between different readers, we provided the detection accuracies among CB cells identified by only one or both readers.

Table 1. Evaluation of the proposed approach over 100 ROI images.

	single pathologist	**both pathologists**	**Non-CB**	**False-positive**
Accuracy	81.4%	89.0%	95.0%	106 cells

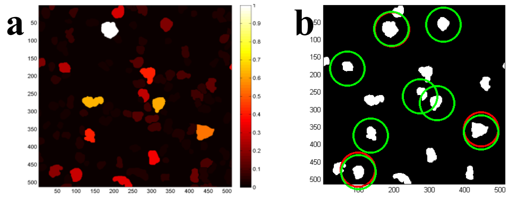

Figure 4. (a) CB likelihood, $_{CB}$, of the sample image given in Figure 2, and (b) detected CBs after adaptive thresholding. Green and red circles indicate ground-truth provided by two pathologists.

5. Conclusions

We developed an image analysis approach for automated detection of CB cells from digitized H&E-stained FL tissue samples. The proposed segmentation approach segments individual cells and detects CBs with high accuracy. Although, the current false positive rate is relatively high, this is an initial detection system designed to have high detection sensitivity (for both CBs and non-CBs). In our future studies, we will address reducing the number of false positive detections by analyzing additional features to differentiate CB cells.

Acknowledgements: The project described was supported by Award Number R01CA134451 from the National Cancer Institute.

References

[1] E. S. Jaffe, et al., Pathology and genetics: Tumours of haematopoietic and lymphoid tissues, IRAC Press, 2001.

[2] S. J. Horning, Something old, something few, something subjective, *J of Clinical Oncology*, 21(1): pp. 1-2, 2003.

[3] A. E. Martinez, et al., Grading of follicular lymphoma: Comparison of routine histology, *Arch of Pathol Lab Med,* 131(7): pp. 1084-1088, 2007.

[4] J. Kong, et al., Computer aided evaluation of neuroblastoma on whole-slide histology images, *Pattern Recognition*, 42(6): pp. 1080-1092, 2009.

[5] O. Sertel, et al., Histopathological image analysis using model-based intermediate representations and color texture, *J. of Signal Proc. Sys.*, 55(1): pp. 169-183, 2009.

[6] D. Comaniciu and P. Meer, Distribution free decomposition of multivariate data, *IEEE Trans. on PAMI*, 2(1): pp. 22-30, 1999.

[7] D. Comaniciu and P. Meer, Mean shift: A robust approach toward feature space analysis, *IEEE Trans. on PAMI*, 24(5): pp. 603-619, 2002.