Distribution Fitting-based Pixel Labeling for Histology Image Segmentation

Lei He*, L. Rodney Long, Sameer Antani, George Thoma
National Library of Medicine, National Institutes of Health
8600 Rockville Pike, Bethesda, MD, USA 20894

ABSTRACT

This paper presents a new pixel labeling algorithm for complex histology image segmentation. For each image pixel, a Gaussian mixture model is applied to estimate its neighborhood intensity distributions. With this local distribution fitting, a set of pixels having a full set of source classes (e.g. nuclei, stroma, connective tissue, and background) in their neighborhoods are identified as the *seeds* for pixel labeling. A seed pixel is labeled by measuring its intensity distance to each of its neighborhood distributions, and the one with the shortest distance is selected to label the seed. For non-seed pixels, we propose two different labeling schemes: *global voting* and *local clustering*. In global voting each seed classifies a non-seed pixel into one of the seed's local distributions, i.e., it casts one vote; the final label for the non-seed pixel is the class which gets the most votes, across all the seeds. In local clustering, each non-seed pixel is labeled by one of its own neighborhood distributions. Because the local distributions in a non-seed pixel neighborhood do not necessarily correspond to distinct source classes (i.e., two or more local distributions may be produced by the same source class), we first identify the "true" source class of each local distribution by using the source classes of the seed pixels and a minimum distance criterion to determine the closest source class. The pixel can then be labeled as belonging to this class. With both labeling schemes, experiments on a set of uterine cervix histology images show encouraging performance of our algorithm when compared with traditional multithresholding and *K*-means clustering, as well as state-of-the-art mean shift clustering, multiphase active contours, and Markov random field-based algorithms.

Keywords: Image segmentation, labeling, histology, local distribution fitting

1. INTRODUCTION

Histology [1, 2] is the study of the microscopic anatomy of cells and tissues of organisms. Histology analysis is performed by examining a thin slice of tissue under an optical or electron microscope. Such tissue samples are usually produced after a sequence of technical procedures, including fixation, dehydration, clearing, infiltration, embedding, sectioning, and staining. In practice, histology image use encompasses diverse modalities from various imaging acquisition technologies [3], based on which manual or automated analysis can be conducted by histopathologists and clinicians.

In present research, histology image interpretation is regarded as the gold standard for clinical diagnosis of cancers and identification of prognostic and therapeutic targets. Histopathologists or clinicians visually examine the regularities of cell shapes and tissue distributions and make diagnostic decisions about cancer presence and degree of malignancy. At the present time, manual analysis of histology continues to be the primary instrument for identifying cancerous tissues, and depends heavily on the expertise and experience of histopathologists. These manual methods have the disadvantages of (a) being very time consuming for such high throughput and high content screening, and (b) lacking consistency and reproducibility in grading; intra- and inter-observation variations remains a serious issue. To attempt to address these problems, computer assisted diagnosis (CAD) systems, which provide rapid and consistent diagnostic results, have been developed for automated histology image analysis. Computer aided methods have been employed for numerous cancer detection and classification applications, such as prostate [4], breast [5], cervix [6], and lung [7] cancer detection and grading; neuroblastoma categorization [8], and follicular lymphoma grading [9].

A typical CAD system consists of a sequence of image processing and machine learning modules, such as image preprocessing, segmentation, feature extraction and dimensionality reduction, disease detection and grading, and post-processing. We briefly describe these steps in the following, image preprocessing reduces the input image noise and enhances features of interest for easier analysis in later modules. Specifically for the high-throughput and high-content tissue screening, multi-scale decomposition may be applied to reduce the computational cost, i.e., low resolution images can be analyzed first to roughly locate the regions of interest (ROI), which will be the focus of higher resolution image analysis. Image segmentation extracts the target objects or regions for feature extraction. Traditional image segmentation methods [10, 11] include edge detection, thresholding, region growing, and *K*-means clustering, which usually requires post-processing such as edge linking or morphological operations to produce continuous and closed boundaries. Recently more advanced approaches such as active contours [12, 13, 20], Markov random field (MRF) models [11, 14], and mean shift clustering [15] have been proposed with promising performance. After segmentation, a variety of image features can be computed from the extracted regions, including morphometrics [4, 6, 8] with object size and shape (e.g. compactness and regularity), graph-based features [5, 6] (e.g. Voronoi diagram and Delaunay triangulation), intensity and color features (e.g. statistics in different color spaces [4]), as well as texture features [5, 7, 9] (e.g. Haralick entropy, Gabor filter, power spectrum, co-occurrence matrices, and wavelets). In addition to these spatial domain features, many features can also be extracted from other transformed spaces, e.g. frequency (Fourier) space and wavelet transformation [16]. In practice, the number of extracted features can be prohibitive for current CAD systems. Therefore, feature dimensionality reduction (DR) techniques [17] are needed to select the most discriminative ones for feasible analysis. The commonly used DR tools include both linear and nonlinear techniques. Linear techniques such as principal component analysis, linear discriminant analysis, and multidimensional scaling are used in cases of points which are linearly separable in the feature space. Nonlinear DR techniques such as spectral clustering, isometric mapping, locally linear embedding, and Laplacian eigenmaps do not assume Euclidean relationship among feature points. These techniques are more suitable for inherently nonlinear biomedical structures. Finally, supervised classification algorithms [11] (e.g. support vector machine and neural network) can be used to classify these simplified feature vectors in order to identify diseased tissues by comparing the input image features with pre-derived training sample features. In certain applications, post-processing may be required to derive high level knowledge from the CAD system results. For example, the extracted object shapes may be specifically indexed for advanced applications like image retrieval. Similarly, image analysis results may be applied for image annotation and information fusion. Note that the sequential order of the functional steps described above may vary in practical applications. For example, texture image segmentation requires that texture features should be computed before segmentation. Also, some steps may be omitted in particular systems, and other application-specific modules which we have not discussed, may be included.

In this paper, we focus on automated image segmentation in a CAD system. Specifically, we propose a new pixel labeling algorithm to extract objects/regions of different classes (e.g. nuclei, stroma, cytoplasm, blood cells, and background) from histology images with complex distributions. With *local distribution fitting techniques*, we classify each pixel according to its similarity to either a set of predetermined seeds (called *global voting*) or its neighborhood (called *local clustering*). The paper is organized as follows. After a brief review of image segmentation approaches in Section 2, we present our proposed approach in Section 3. Section 4 shows experimental results, including the comparison with traditional multithresholding and *K*-Means clustering, as well as state-of-the-art multiphase active contours, mean shift clustering, and MRF-based methods. We present our conclusions in Section 5.

2. BACKGROUND

Early segmentation methods [10, 11] include thresholding, edge detection, region growing, and *K*-means clustering. Thresholding approaches search for a value (intensity threshold) to separate objects from background. The threshold is usually identified to satisfy some constraints or to optimize certain objective functions. For example, the commonly used Otsu's method [10] finds the threshold to maximize the between-class variance. In the case of a histology image *I* (size *X* × *Y*) with *K* object classes (s_1, s_2, …, s_K), Otsu's method finds the thresholds that maximizes the between-class variance

$$\sigma_B^2 = \sum_{k=1}^{K} P_k\,(\mu_k - \mu_G)\,, \tag{1}$$

where $P_k = \sum_{l\in s_k} p_l$ and p_l is the normalized histogram (probability) of intensity *l*, i.e., $p_l = n_l/XY$ and n_l is the number of pixels with intensity *l*. μ_k is the current mean of s_k, $\mu_k = \frac{1}{P_k}\sum_{i\in s_k} lp_l$, and μ_G is the whole image intensity mean. The *K* classes are separated by *K*-1 thresholds that maximize $\sigma_B{}^2$. Edge detection applies spatial filters (e.g. Canny and Sobel filters) to determine the border among objects and background. Region growing [10] groups pixels with similar features (e.g. intensity or texture) into connected areas, each of which is regarded as homogenous or smooth according to predefined feature similarities. *K*-means clustering classifies an image point into one of the *K* clusters by minimizing the objective function:

$$\sum_{i=1}^{K}\sum_{\mathbf{x}\in C_i}| I_{\mathbf{x}} - \mu_i |^2 , \tag{2}$$

where $I_{\mathbf{x}}$ is the intensity of a pixel $\mathbf{x}\in\Re^2$, in the class C_i, and μ_i is the current mean of C_i. Both thresholding and *K*-means clustering need post-processing operations to remove noise (spurious) edges and produce continuous object boundaries.

Typical difficulties in image segmentation include noise, low intensity contrast with weak edges, and intensity inhomogeneity [18], which pose significant challenges to traditional segmentation methods. To address these difficulties, more advanced methods, such as the mean shift clustering, MRF-based pixel labeling methods, and active contour models, have been proposed for segmentation with promising results. Unlike the *K*-means clustering, the mean shift algorithm [15] does not assume prior knowledge of the number of clusters. For image segmentation, the image points in a *d*-dimensional (*d*=3 for color image) feature space can be characterized by a probability density function where dense regions correspond to the local maxima (modes) of the underlying distribution. Image points associated with the same mode (by a gradient ascent procedure) are grouped into one cluster. MRF approaches use a Bayesian framework to map a random field (with Gibbs distribution) to an image in which each pixel is characterized by a random variable with all possible class labels. The result is that the segmentation process is formulated as pixel labeling by maximizing the posterior probability of the labeled configuration, given the observation. In practice both deterministic (e.g. iterated conditional modes (ICM) [11] and graph cuts [14]) and stochastic algorithms (Metropolis and Gibbs sampling [11]) are used for the maximum a posterior estimation.

Compared with above segmentation techniques, the active contour models can achieve subpixel accuracy and provide closed and smooth contours/surfaces, which become an increasingly important tool for microscopy image segmentation [19]. For example, the well-known Chan-Vese model (CV) [12] assumes homogeneous object and background regions with distinct intensity means. Given a gray scale image I_0: $\Omega \subset \Re^2 \rightarrow \Re$, the CV energy functional is defined as:

$$E(c_1,c_2,\phi) = \int_\Omega (I_0 - c_1)^2 H(\phi)d\mathrm{x} + \int_\Omega (I_0 - c_2)^2 (1 - H(\phi))d\mathrm{x} + \nu\int_\Omega | \nabla H(\phi) | \, d\mathrm{x} , \tag{3}$$

where ν>0 is a constant. c_1 and c_2 are two global constants representing the intensity means of the two regions, i.e., background and objects. H is the Heaviside step function and ϕ represents the level set function. Eq. (3) handles images with two different regions. To segment histology images with multiple object classes, Eq. (3) has to be extended to multiphase level sets [20]. The extended energy functional is:

$$\begin{aligned} E(\mathrm{c},\Phi) = &\int_\Omega (I - \mathrm{c}_{11})^2 H(\phi_1)H(\phi_2)dx + \int_\Omega (I - \mathrm{c}_{10})^2 H(\phi_1)(1 - H(\phi_2))d\mathrm{x} + \int_\Omega (I - \mathrm{c}_{01})^2 (1 - H(\phi_1))H(\phi_2)dx \\ &+ \int_\Omega (I - \mathrm{c}_{00})^2 (1 - \mathrm{H}(\phi_1))(1 - H(\phi_2))d\mathrm{x} + \nu\int_\Omega | \nabla H(\phi_1) | \, dx + \nu\int_\Omega | \nabla H(\phi_2) | \, dx \end{aligned} , \tag{4}$$

where c = (c_{00}, c_{01}, c_{10}, c_{11}) represents the average values of four image regions produced by two level sets Φ = (ϕ_1, ϕ_2). By calculus of variations, the level set evolution equation can be derived as:

$$\begin{aligned} \frac{\partial\phi_1}{\partial t} &= \delta(\phi_1)\{\nu\,\mathrm{div}(\frac{\nabla\phi_1}{|\nabla\phi_1|}) - ((I_0 - \mathrm{c}_{11})^2 - (I_0 - \mathrm{c}_{01})^2)H(\phi_2) - ((I_0 - \mathrm{c}_{10})^2 - (I_0 - \mathrm{c}_{00})^2)(1 - H(\phi_2))\}, \\ \frac{\partial\phi_2}{\partial t} &= \delta(\phi_2)\{\nu\,\mathrm{div}(\frac{\nabla\phi_2}{|\nabla\phi_2|}) - ((I_0 - \mathrm{c}_{11})^2 - (I_0 - \mathrm{c}_{10})^2)H(\phi_1) - ((I_0 - \mathrm{c}_{01})^2 - (I_0 - \mathrm{c}_{00})^2)(1 - H(\phi_1))\} \end{aligned} \tag{5}$$

In Eq. (5), δ is the Dirac function.

With a different framework for extracting multiple objects, Samson's image classification model [13] applies a group of level sets, (ϕ_1,…, ϕ_K), to divide the input image into K regions, each of which corresponds to the interior of a level set, i.e., ϕ_i>0.

$$E(\phi_1,...,\phi_K)=\sum_{i=1}^{K}e_i\int_\Omega H(\phi_i)\frac{(I_0-\mu_i)^2}{\sigma_i^2}dx+\sum_{i=1}^{K}\gamma_i\int_\Omega g(I_0)\delta(\phi_i)\,|\nabla\phi_i\,|\,dx+\frac{\lambda}{2}\int_\Omega(\sum_{i=1}^{K}H(\phi_i)-1)^2\,dx\,, \tag{6}$$

where e_i>0, γ_i>0, and λ>0 are constants to balance the terms. $g(I_0)=\dfrac{1}{1+|\nabla G_\sigma * I_0|^2}$ is a monotonically decreasing function, which deforms contours towards edges. Briefly, the first term in Eq. (6) ensure a homogeneous region within each level set, which can be fitted by a Gaussian distribution with pre-estimated mean μ_i and variance σ_i^2. The second term prefers a smooth curve at edges, and the third term prevents overlapping level sets. The level set evolution equations are:

$$\frac{\partial\phi_i}{\partial t}=-\delta(\phi_i)\{e_i\frac{(I_0-\mu_i)^2}{\sigma_i^2}-\gamma_i g(I_0)div(\frac{\nabla\phi_i}{|\nabla\phi_i\,|})-\frac{\nabla g\nabla\phi_i}{|\nabla\phi_i\,|}+\lambda(\sum_{i=1}^{K}H(\phi_i)-1)\}, \tag{7}$$

In this paper we compare our pixel labeling algorithm with traditional multithresholding and K-means clustering, as well as the advanced mean shift clustering, MRF-based methods, and active contours (the multiphase CV model (Eq. (5)) and Samson's model (Eq. (7)).

3. PROPOSED APPROACH

Histology images consist of a large quantity of objects of interest (usually cells and prominent cell structures, such as nuclei) widely distributed in the images and surrounded by different neighboring tissues (for example, in the cervix, epithelium and stroma). We propose a new pixel labeling algorithm for such complex histology image segmentation. We assume that an image pixel always belongs to a class (distribution) in its own neighborhood, and that these local distributions can be modeled by Gaussians with different means and variances, i.e., by a Gaussian mixture model (GMM).

3.1 Gaussian mixture model

Mixture models [21] are widely used to approximate complicated distributions with the output coming from one of a group of "hidden" sources (e.g. objects and background in an image), which provides a general framework to characterize heterogeneity. In this paper, we choose mixture models for our specific application of histology image segmentation, where we expect multiple classes of objects widely distributed in the image. In statistics, a mixture model is usually defined as a probability distribution that is a convex combination of several independent components, where the components are themselves characterized by different probability distributions. Given an output, the goal is to estimate from which source (measured by probabilities) the output is generated, as well as the parameters describing the source component distributions, e.g. means and variances of a GMM. With a set of N samples (image points) from n-dimensional space, X = {$\mathbf{x}_1$, …, $\mathbf{x}_j$, …, $\mathbf{x}_N$}, in which each sample is drawn from one of M Gaussian components, a GMM can be denoted as:

$$p(X\,|\,\Theta)=\sum_{i=1}^{M}\alpha_i\,p_i(X\,|\,\theta_i)\,, \tag{8}$$

where the parameters are Θ={α_1 ,…, α_M, θ_1 ,…, θ_M} such that $\sum_{i=1}^{M}\alpha_i=1$ and α_i refers to the prior probability of each component; $\theta_i=(\mu_i,\Sigma_i)$, μ_i is the mean and $\sum_i$ is the covariance matrix, i=1, … M. Let y_j, j=1, … N, denote which Gaussian $\mathbf{x}_j$ is drawn from, the probability of $\mathbf{x}_j$ coming from the i-th Gaussian is:

$$P(\mathbf{x}_j \mid y_j = i, \theta_i) = \frac{\exp\left(-\frac{1}{2}(\mathbf{x}_j - \mu_i)^T \Sigma_i^{-1} (\mathbf{x}_j - \mu_i)\right)}{(2\pi)^{n/2} \mid \Sigma_i \mid^{1/2}}, \tag{9}$$

The task is to estimate the hidden distributions given the data, i.e., to estimate the unknown parameters Θ which maximize Eq. (9). The GMM parameters can be estimated by the expectation-maximization (EM) algorithm [22], which repeats the E-step and M-step until convergence. The E-step is to calculate the expectation of which Gaussian is used, conditioned on the observations (X), using the estimated prior probability of each distribution ($p(y_j=i|\theta_i)$) and current parameter values (Θ_t),

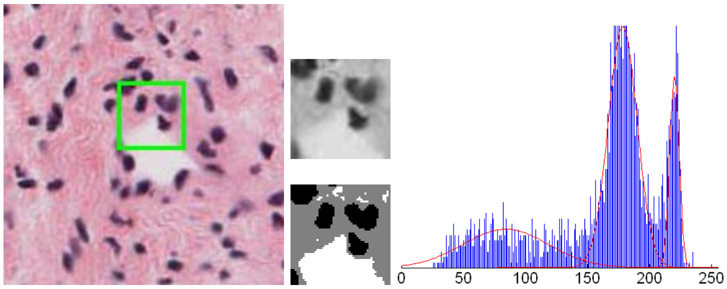
$$p(y_j = i \mid \mathbf{x}_j, \Theta_t) = \frac{p(\mathbf{x}_j \mid y_j = i, \Theta_t)\, p(y_j = i \mid \Theta_t)}{\sum_{k=1}^{M} p(\mathbf{x}_j \mid y_j = k, \Theta_t)\, p(y_j = k \mid \Theta_t)}, \tag{10}$$

Given the E-step estimation of unknown variables ($\mathbf{y} = \{y_1,\ldots,y_N\}$, $y_j=1,\ldots,M$), the M-step estimates the distribution parameters (Θ) and the prior probability of each distribution, which maximize the data likelihood as

$$Q(\Theta, \Theta_t) = E_y\left[\log \prod_{j=1}^{N} p(\mathrm{x}_j, \mathrm{y} \mid \Theta) \middle| \mathrm{x}_j\right] = \sum_{j=1}^{N}\sum_{i=1}^{M} p(y_j = i \mid \mathrm{x}_j, \Theta_t) \log(p(\mathrm{x}_j \mid y_j = i, \Theta)\, p(y_j = i \mid \Theta)), \tag{11}$$

where the log-likelihood is used for easier numerical implementation. With gradient ascent approach, we can update the parameters and the prior probabilities as:

$$\mu_i = \frac{\sum_{j=1}^{N} p(y_j = i \mid \mathbf{x}_j, \Theta_t)\mathbf{x}_j}{\sum_{j=1}^{N} p(y_j = i \mid \mathbf{x}_j, \Theta_t)}, \text{ and } \Sigma_i = \frac{\sum_{j=1}^{N} p(y_j = i \mid \mathbf{x}_j, \Theta_t)(\mathbf{x}_j - \mu_i)(\mathbf{x}_j - \mu_i)^T}{\sum_{j=1}^{N} p(y_j = i \mid \mathbf{x}_j, \Theta_t)} \tag{12}$$



(a) (b) (c)

Figure 1. Local GMM distribution estimation example. (a) Original image with a region of interest. (b) Top: gray version of ROI in (a); Bottom: classes estimated by GMM. (c) Local intensity distribution fitting of (b) by GMM.

These updated parameters then become the input for next E-step, and the convergence to a local maximum of the EM algorithm is guaranteed [22]. Using the EM algorithm, Figure 1 shows an example of local GMM-based distribution estimation in a small region (see the green rectangle). The left image in Figure 1 is an input image with three target classes of nuclei, cytoplasm, and background, which is cropped (size 200×200) from a large size cervix histology image (size 72360×41788). The middle top image is the gray scale version of the selected region. The middle bottom image is the segmentation result based on the estimated distribution, i.e., each pixel is grouped to the cluster (distribution) to which it has the closest distance. The blue lines in the right image correspond to the intensity histogram of the region. The estimated Gaussian distributions of the objects and background are illustrated as the red curves. It can be seen that the estimated distributions match well with the real intensity histogram, which show the suitability to use the GMM for objects and background distribution estimation.

3.2 Pixel labeling algorithm

For our algorithm, we define the neighborhood of an image pixel as an $r \times r$ region centered on the pixel, where the size r is an algorithm parameter (e.g. $r = 11$). Then for each pixel, we estimate its neighborhood intensity distributions with a Gaussian mixture model. Given the intensity value of a pixel and the estimated GMM for the pixel's neighborhood, the goal is to estimate from which source class (measured by probabilities) the pixel value is generated. Therefore, for each image region under the scanning window, the local intensity distribution is estimated by a local GMM with a prior number (K) of different target classes in the image. Since we use a fixed number K of distributions for each GMM, the computed local distributions may not truly represent different source classes. A pixel within a large homogeneous region, for example, will actually have only one source class in its neighborhood. To identify the source classes, we apply a confidence test to each GMM to attempt to distinguish between pixels with K "true" neighborhood classes, and those with fewer than K. The test is implemented by evaluating the distances among different distributions, i.e., two distributions are considered to represent two source classes only when their distance is larger than a predefined threshold (T). Bhattacharyya distance is applied to measure the distance of two Gaussian distributions:

$$d_B(N(\mu_1,\sigma_1),N(\mu_2,\sigma_2)) = \frac{(\mu_1-\mu_2)^2}{4(\sigma_1+\sigma_2)} + \frac{1}{2}\log\frac{|\sigma_1+\sigma_2|}{2\sqrt{\sigma_1\sigma_2}}, \tag{13}$$

We call the pixels with all K different source classes in their neighborhoods "seed" pixels. We assign each seed a source class label by comparing its intensity distance to its neighborhood Gaussian distributions, i.e.,

$$d_i(\mathbf{x}_j) = \frac{1}{(2\pi)^{n/2}\,|\Sigma_i|^{1/2}}\exp\left(-\frac{1}{2}(\mathbf{x}_j - \mu_i)^T\,\Sigma_i^{-1}(\mathbf{x}_j - \mu_i)\right), \tag{14}$$

This measures the probability of the pixel $\mathbf{x}_j$ being generated by the ith Gaussian distribution in its neighborhood.

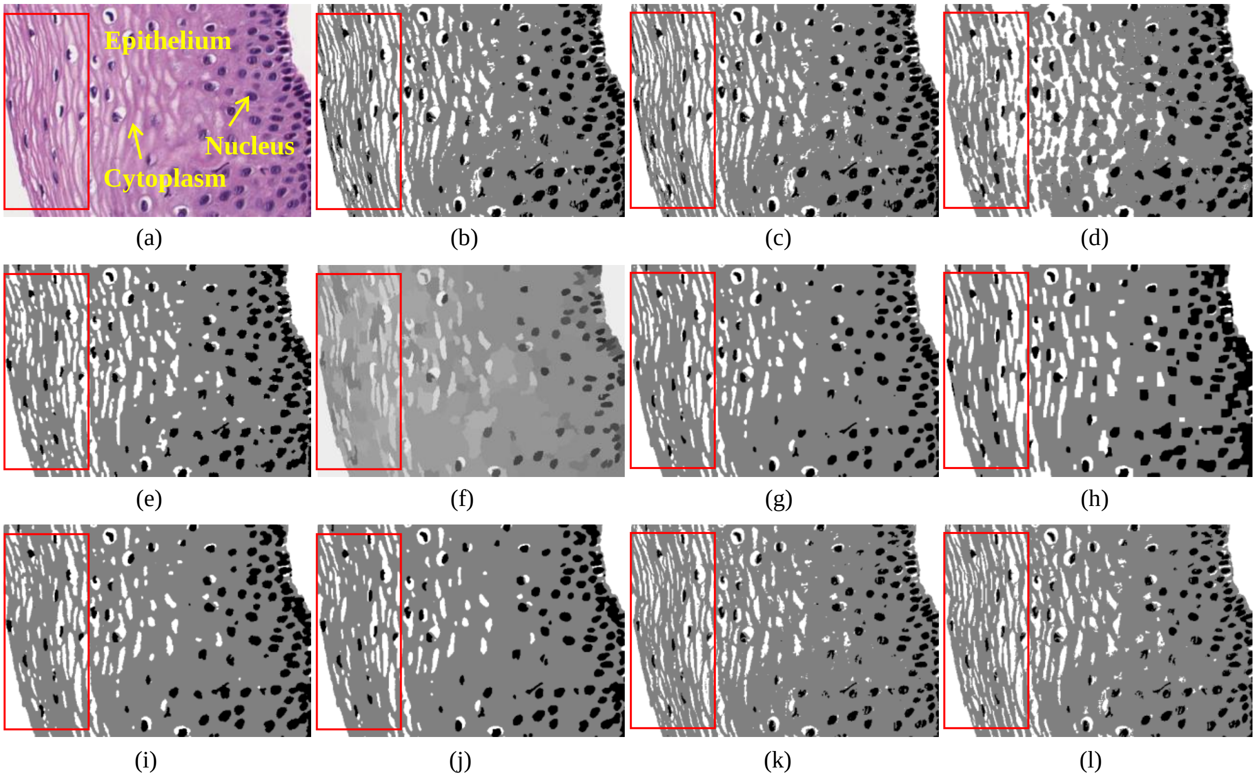
To label non-seed pixels, we propose two different schemes: *global voting* and *local clustering*. In the global voting method, given a non-seed pixel, each seed pixel "votes" for one of its K neighborhood classes as being the most likely class using Eq. (14), viz., the class with the smallest distance to the non-seed pixel is used to label that pixel. The final label of the non-seed pixel is determined by tallying the votes cast by all the seeds. In practice, the larger the number of the seeds, the more confident the voting and thus the more accurate the segmentation. On the other hand, a large number of seeds always introduces a high computational cost. Thus the threshold T is chosen to balance the performance and the cost. In our experiments, we choose the threshold that produces about 100 seeds.

In the local clustering method, a non-seed pixel is labeled directly with one of its own neighborhood distributions. The estimated neighborhood distributions of a non-seed pixel may not correspond to distinct source classes (i.e., two neighborhood distributions may correspond to the same source class). Because of this, the "true" source classes of the distributions have to be determined first by comparing them with those of the seeds by Eq. (13). Again, the local distribution classes of each non-seed pixel are computed by voting of all the seeds. In this manner the non-seed pixels can be finally labeled by the identified neighborhood clusters.

4. EXPERIMENTS

This section presents the experiments on a set of three cervix histology images with complex distributions. We compare our model to nine major image segmentation methods, including traditional multithresholding [10] and K-means clustering [11], as well as the more contemporary mean shift clustering [15], multiphase level set models [13, 20], and MRF-based labeling methods (including both deterministic algorithms: iterated conditional modes (ICM) [11] and graph cut [14], and nondeterministic algorithms: Metropolis algorithm [11] and Gibbs sampling [11]). Overall our algorithm obtains segmented regions that are visually comparable to, if not better than, the regions obtained by the methods listed above. For the active contour methods, multiple uniformly distributed small rectangles were used as the initial contours.

In Figure 2(a), a cervix epithelium tissue image is shown with three target classes: nuclei, cytoplasm, and background. For this simple example, all methods can obtain results that might be considered acceptable; however, there are observable differences, some of which we note here. Because the mean shift clustering does not use the prior number of source classes, it usually produces more (or fewer) classes than the expected number (Figure 2(f)), which thus requires post-processing such as region merging or splitting for final results. We can also observe that minor texture details are missed in most of the results, except the traditional multithresholding (Figure 2(b)), K-means clustering (Figure 2(c)) and our pixel labeling algorithm (Figures 2(k), (l)), as shown by the red rectangles. In particular, MRF-based labeling approaches (Figures 2(g)-(j)) show loss of details by under-segmentation of the cytoplasm areas. For another example, Figure 3(a) shows an image of cervix stroma with three source classes: nuclei, connective tissue, and background. While most methods successfully extract the nuclei, none can accurately segment the connective tissue from the background due to the rather noisy and inhomogeneous content. MRF-based segmentation methods (Figures 3(g), (i), and (j)) and our global voting-based pixel labeling algorithm (Figure 3(k)) obtain better results than others. Nevertheless, MRF-based segmentation results show oversegmented nuclei, and our results include some background noise in the tissue constituents. Figure 4(a) consists of one more class than Figure 2: blood cells. Again, MRF-based methods (Figures 4(g), (i)) and our algorithm (Figures 4(k), (l)) obtain better results. Other methods produce larger segmentation error on separating the blood cells from the cytoplasm at the lower left part.

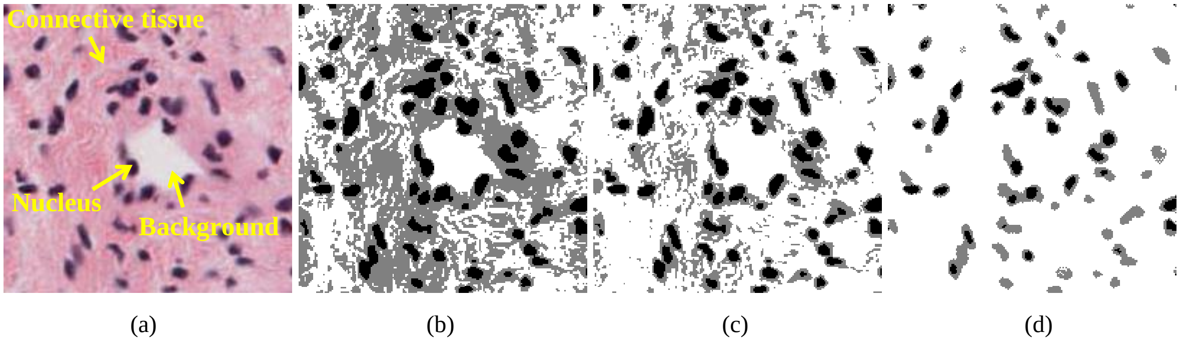


(a) (b) (c) (d)

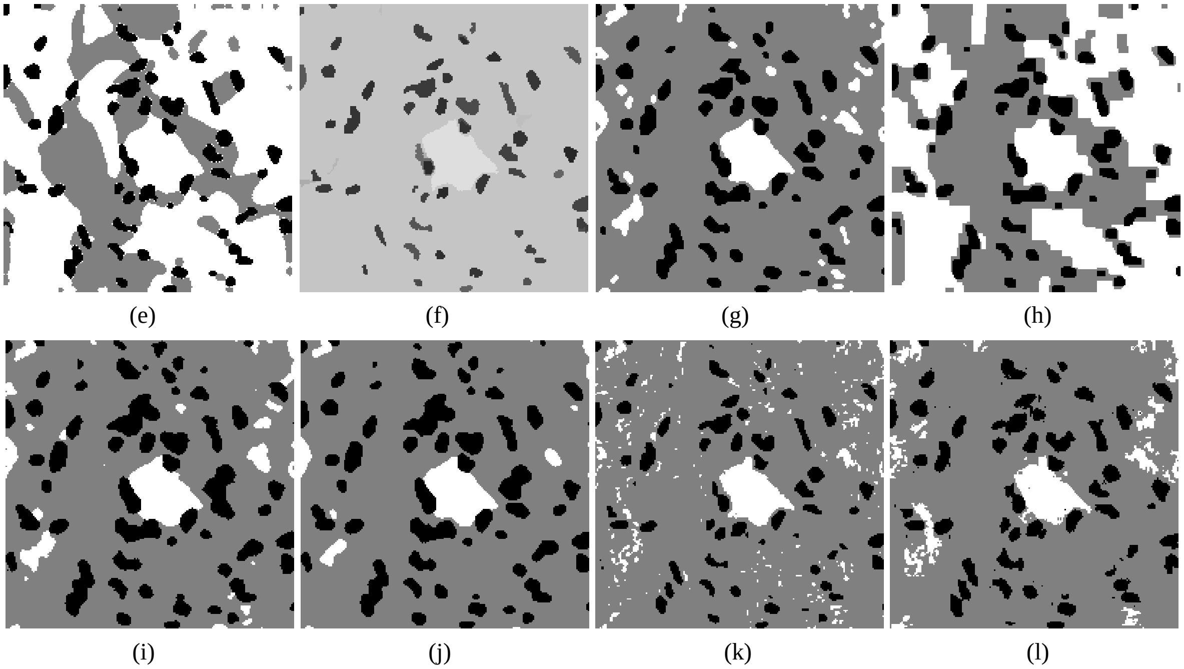
(e) (f) (g) (h)

(i) (j) (k) (l)

Figure 2. Cervix epithelium histology image segmentation results (a) Original histology image with three classes. (b) Multithresholding [10]. (c) *K*-means clustering [11]. (d) Multiphase level sets [20]. (e) Samson's model [13]. (f) Mean shift clustering [15]. (g) MRF by ICM [11]. (h) MRF by graph cuts [14]. (i) MRF by Metropolis algorithm [11]. (j) MRF by Gibbs sampling [11]. (k) Our method: pixel labeling by global voting. (l) Our method: pixel labeling by local clustering.



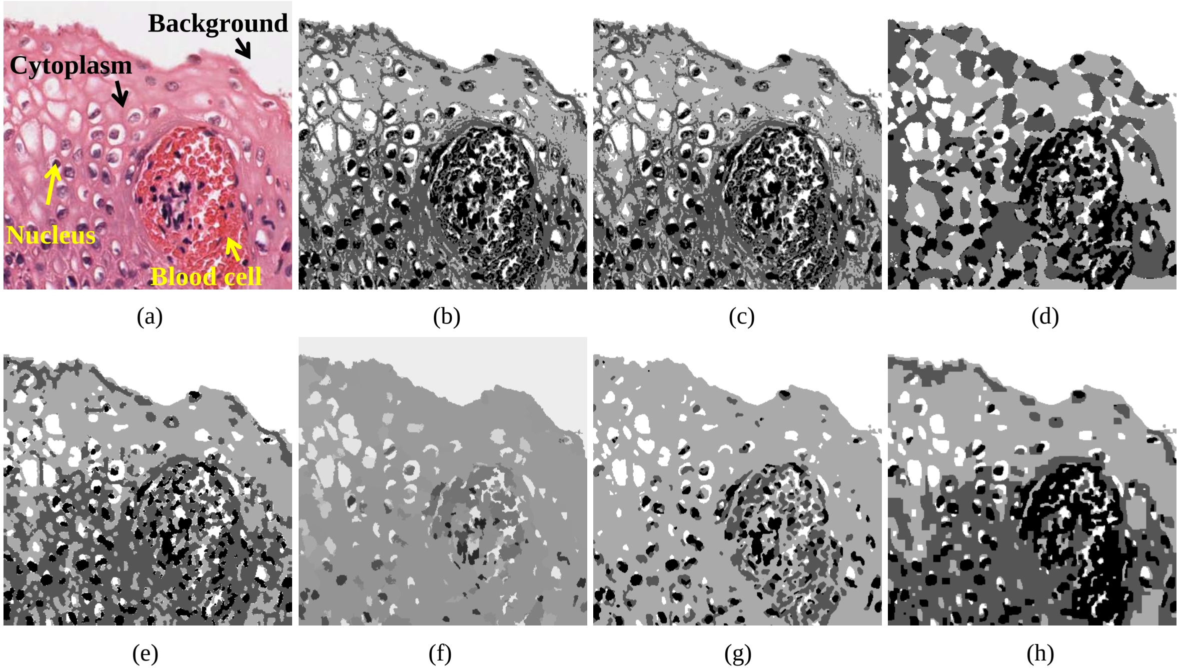
(a) (b) (c) (d)



(e) (f) (g) (h)

(i) (j) (k) (l)

Figure 3. Cervix stroma histology image segmentation results (a) Original histology image with three classes. (b) Multithresholding. (c) *K*-means clustering. (d) Multiphase level sets. (e) Samson's model. (f) Mean shift clustering. (g) MRF by ICM. (h) MRF by graph cuts. (i) MRF by Metropolis algorithm. (j) MRF by Gibbs sampling. (k) Our method: pixel labeling by global voting. (l) Our method: pixel labeling by local clustering.



(a) (b) (c) (d)

(e) (f) (g) (h)

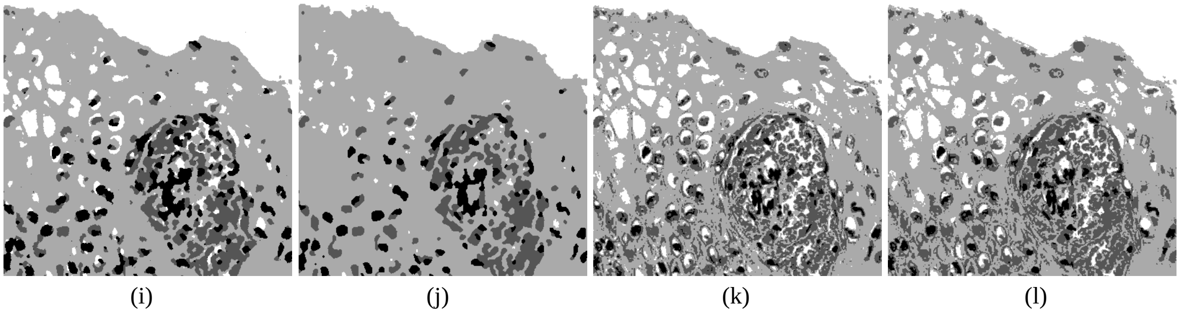


Figure 4. Cervix stroma histology image segmentation results (a) Original histology image with three classes. (b) Multithresholding. (c) *K*-means clustering. (d) Multiphase level sets. (e) Samson's model. (f) Mean shift clustering. (g) MRF by ICM. (h) MRF by graph cuts. (i) MRF by Metropolis algorithm. (j) MRF by Gibbs sampling. (k) Our method: pixel labeling by global voting. (l) Our method: pixel labeling by local clustering.

5. CONCLUSION

This paper presents a new pixel labeling algorithm for complex histology image segmentation. We characterize local image intensity distributions by Gaussian mixture models, which are then used to label pixels as members of one of *K* source classes in the image. We rely on the concept of "seed" pixels, i.e., the image pixels where neighborhood pixel distributions give strong evidence that we have the full *K* classes represented in the neighborhood. For each seed pixel, we use the distribution parameters of these neighborhood classes to determine the seed label. We propose two different schemes for labeling non-seed pixels: global voting and local clustering. Compared with nine traditional and state-of-the-art methods, our model provides a simple and flexible framework for histology image segmentation; the only parameters are local neighborhood size (r) and the distribution distance threshold (T). In these experiments, our results appear favorable with respect to visual inspection, for separation of the major tissue classes in the image. We are investigating additional techniques, such as nondeterministic label propagation [23] to further improve the performance. Our planned future work also includes feature extraction in the classification for the segmented objects and regions, toward the goal of computer assisted cancer detection and malignancy level grading, e.g. cervical intraepithelial neoplasia (CIN) grading for cervix histology images.

ACKNOWLEDGEMENT

This research was supported by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM), and Lister Hill National Center for Biomedical Communications (LHNCBC).

REFERENCES

[1] Ross, M., Kaye, G. I. and Pawlina, W., [Histology: a Text and Atlas], Lippincott Williams & Wilkins, 4th Ed. (2002).

[2] Jungueira, L. and Carneiro, J., [Basic Histology: Text & Atlas], McGraw-Hill Medical, 11th Ed. (2005).

[3] Murphy, D. B., [Fundamentals of Light Microscopy and Electronic Imaging], Wiley-Liss, (2001).

[4] Monaco, J., Tomaszewski, J., Feldman, M., Mehdi, M., Mousavi, P., Boag, A., Davidson, C., Abolmaesumi, C. and Madabhushi, A., "Probabilistic pair-wise Markov models: application to prostate cancer detection," Proc. SPIE Medical Imaging 7260, (2009).

[5] Doyle, S., Agner, S., Madabhushi, A., Feldman, M. and Tomaszewski, J., "Automated grading of breast cancer histopathology using spectral clustering with textural and architectural image features," Proc. IEEE International Symposium on Biomedical Imaging, 496-499 (2008).

[6] Keenan, S. J., Diamond, J., McCluggage, W. G., Bharucha, H., Thompson, D., Bartels, P. H. and Hamilton, P. W., "An automated machine vision system for the histological grading of cervical intraepithelial neoplasia (CIN)," Journal of Pathology 192, 351-362 (2000).

[7] Schmid, K., Angerstein, N., Geleff, S. and Gschwendtner, A., “Quantitative nuclear texture features analysis confirms WHO classification 2004 for lung carcinomas,” Modern Pathology 19, 453-459 (2006).

[8] Kong, J., Sertel, O., Shimada, H., Boyer, K. L., Saltz, J. H. and Gurcan, M. N., “Computer-aided evaluation of neuroblastoma on whole-slide histology images: classifying grade of neuroblastic differentiation,” Pattern Recognition 42, 1080-1092 (2009).

[9] Sertel, O., Kong, J., Lozanski, G., Catalyurek, U., Saltz, J. and Gurcan, M. N. “Computerized microscopic image analysis of follicular lymphoma,” Proc. SPIE Medical Imaging 6915, (2008).

[10] Gonzalez, R. C. and Woods, R. E., [Digital Image Processing], Pearson Prentical Hall, 3rd Ed. (2008).

[11] Bishop, C. M., [Pattern Recognition and Machine Learning], Springer, (2006).

[12] Chan, T. F. and Vese, L. A. “Active contour without edges,” IEEE Trans. on Image Processing 10, 266-277 (2001).

[13] Samson, C., Blanc-Féraud, L., Aubert, G. and Zerubia, J., “A level set model for image classification,” International Journal of Computer Vision 40, 187-197 (2000).

[14] Boykov, Y., Veksler, L. and Zabih, R., “Fast approximate energy minimization via graph cuts,” IEEE Trans. Pattern Analysis and Machine Intelligence17(2), 158-171 (1995).

[15] Comaniciu, D. and Meer, P., “Mean shift: a robust approach towards feature space analysis,” IEEE Trans. Pattern Analysis and Machine Intelligence 24(5), 603-619 (2002).

[16] Orlov, N., Johnston, J., Macura, T., Shamir, L. and Goldberg, I., “Computer vision for microscopy applications,” in [Vision Systems – Segmentation and Pattern Recognition], Obinata, G. and Dutta, A. V. Eds. ARS Press, 221-242 (2007).

[17] Lee, G., Rodriguez, C., Madabhushi, A., “Investigating the efficacy of nonlinear dimensionality reduction schemes in classifying gene and protein expression studies,” IEEE/ACM Trans. Computational Biology and Bioinformatics 5(3), 368–384 (2008).

[18] He, L., Peng, Z., Everding, B., Wang, X., Han, C., Weiss, K. and Wee, W. G., “A comparative study of deformable contour method in medical image segmentation,” Image and Vision Computing 26, 141-163 (2008).

[19] Hafiane, A., Bunyak, F. and Palaniappan, K., “Level set-based histology image segmentation with region-based comparison,” Proc. Microscopic Image Analysis with Applications in Biology, (2008).

[20] Vese, L. A. and Chan, T. F., “A multiphase level set framework for image segmentation using the Mumford and Shah model,” International Journal of Computer Vision 50, 271-293 (2002).

[21] McLachlan, G. J. and Peel, D., [Finite Mixture Models], Wiley, (2000).

[22] Dempster, A., Laird, N. and Rubin, D., “Maximum likelihood from incomplete data via the EM algorithm,” Journal of the Royal Statistical Society - Series B 39, 1-38 (1977).

[23] Zhu, X., [Semi-supervised learning with graphs], doctoral dissertation, Carnegie Mellon University, CMU-LTI-05-192, (2005).