31st Annual International Conference of the IEEE EMBS
Minneapolis, Minnesota, USA, September 2-6, 2009

Extraction of informative cell features by segmentation of densely clustered tissue images

Sonal Kothari, Qaiser Chaudry, and May D. Wang, *Member, IEEE*

Abstract— **This paper presents a fast methodology for the estimation of informative cell features from densely clustered RGB tissue images. The features estimated include nuclei count, nuclei size distribution, nuclei eccentricity (roundness) distribution, nuclei closeness distribution and cluster size distribution. Our methodology is a three step technique. Firstly, we generate a binary nuclei mask from an RGB tissue image by color segmentation. Secondly, we segment nuclei clusters present in the binary mask into individual nuclei by concavity detection and ellipse fitting. Finally, we estimate informative features for every nuclei and their distribution for the complete image. The main focus of our work is the development of a fast and accurate nuclei cluster segmentation technique for densely clustered tissue images. We also developed a simple graphical user interface (GUI) for our application which requires minimal user interaction and can efficiently extract features from nuclei clusters, making it feasible for clinical applications (less than 2 minutes for a 1.9 megapixel tissue image).**

I. Introduction

Cytological features of a tissue image including nuclei count, nuclei size distribution, and nuclei shape distribution are important features for decision making in pathology. They have been cited by various authors for cancer grading, cancer subtype classification, extraction of the malignant portion of a tissue image, and analysis of cell therapy progress [1, 2, 3, 4].

Extraction of these features from images in which nuclei are not clustered is possible by using image segmentation. However, in pathological conditions, nuclei in tissues are mostly clustered, necessitating cluster segmentation. Recent work in cluster segmentation [4, 5, 6] shows problems in segmenting complex clusters or suffers from long processing time because of the complex methodology involved. Previous work suggested a nuclei cluster segmentation technique [7]. In this paper, we 1) suggest improvements to that technique, 2) present computational time analysis for cluster segmentation of different types of images, 3) describe informative features extracted from tissue images using the segmented nuclei, 4) describe the graphical user interface (GUI) developed for this methodology, and 5) evaluate features for a set of tissue images and compare them. The overall block diagram of our method is presented in figure 1.

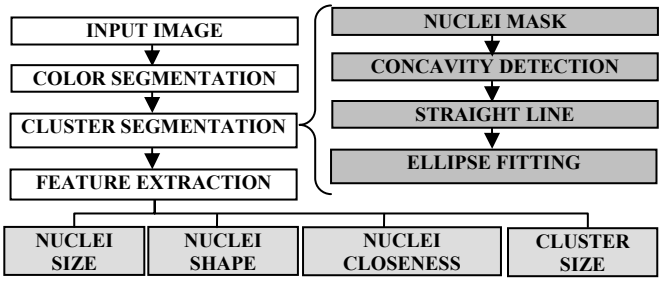

Figure 1: Flow diagram for complete methodology

II. Color Segmentation

Photo micrographs of a stained biopsy tissue section are RGB images with various entities in a tissue slice. These entities include nuclei, glands, cytoplasm and red blood cells appearing as different colors. Therefore, the first step in our algorithm involves color segmentation of the RGB tissue image to obtain a binary mask of the nuclei. With the aim of developing a methodology that may be used for different kinds of stained tissue samples, we have used user-interactive K-means clustering for color segmentation. K-means clustering [8] divides colored pixels of an image into K clusters by minimizing the energy functional E, given by

$$E = \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in \mathrm{S_i}} \left\| \mathbf{x}_j - \mathbf{c}_i \right\|^2 \qquad (1)$$

where E is the sum of the Euclidian distance between pixels $\mathbf{x}_j$ belonging to cluster S_i and its mean $\mathbf{c}_i$ summed for all the clusters. In a user-interactive variation of K-means clustering [9], the user selects approximate seed colors (cluster means) for nuclei and background shades. The GUI provides the user with flexibility to choose the number of nuclei and background shades and then change the means slightly with the sliders in the GUI until a visually suitable mask is obtained. Figure 3(b) is the binary nuclei mask for the RGB tissue image in Figure 3(a).

Manuscript received April 7, 2009. This research has been supported by grants from National Institutes of Health (Bioengineering Research Partnership R01CA108468, P20GM072069, and CCNE U54CA119338), Georgia Cancer Coalition, Hewlett Packard, and Microsoft Research. RCC subtypes data was provided by Dr. Andrew N Young, Emory University. H&N cancer data was provided by Dr. Georgia Z. Chen, Emory University, Atlanta, GA 30332 USA.

Sonal Kothari is with the Department of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA. (phone: 404-385-5059; e-mail: sk9@ gatech.edu).

Qaiser Chaudry is with the Department of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA. (qaiser@gatech.edu).

May D. Wang is with the Department of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA and with the Department of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta, GA 30332 USA (maywang@bme.gatech.edu).

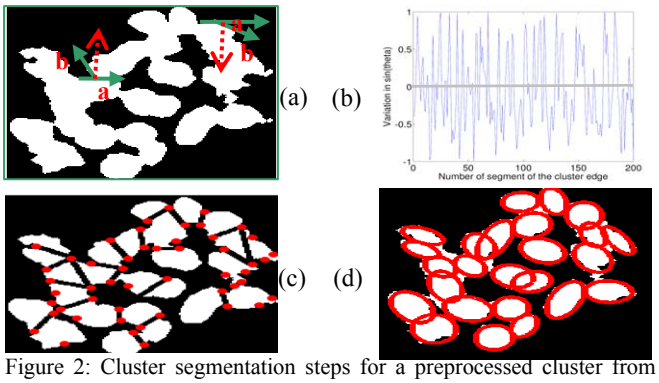

Figure 2: Cluster segmentation steps for a preprocessed cluster from papillary renal cell carcinoma tissue image, (a) nuclei cluster mask with vector **a**, **b** and direction of *sin(θ)* marked at concave and convex edge points, (b) graph depicting variation in *sin(θ)* with segment number of the cluster edge, (c) straight line segmented cluster with concavities marked, (d) ellipse fitting result.

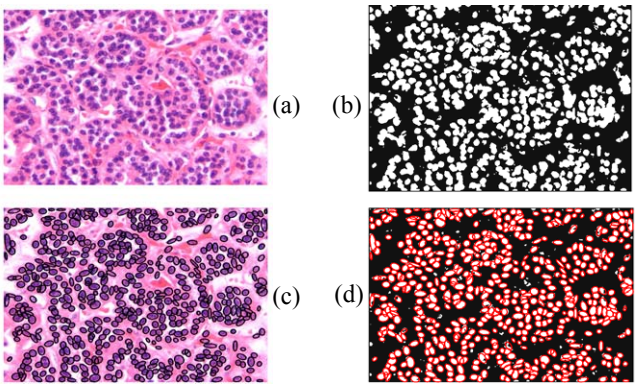

Figure 3: Color and cluster segmentation for an oncocytoma renal cell carcinoma tissue image. (a) RGB image shown in gray scale, (b) binary nuclei mask, (c) individual nuclei marked on RGB tissue image, (d) individual nuclei marked on nuclei mask

III. Cluster Segmentation

The nuclei mask obtained by color segmentation usually has clusters of nuclei. Therefore, the next step involves segmentation of these clusters into individual nuclei. Previous work suggested a methodology for cluster segmentation using concavity detection and ellipse fitting [7]. The methodology used in this paper is similar with some improvements.

A. Preprocessing

The preprocessing steps that we used are the same as those mentioned in previous work [7]. The nuclei mask obtained from color segmentation typically has holes in the nuclei mask as well as noise. Our methodology is edge-based and holes in the mask can lead to false segmentation. Therefore, we fill the holes using morphological reconstruction [10]. Noise in the nuclei mask is due to misclassification by K-means or due to the presence of small portions of nuclei shades in the background of a stained tissue sample. Noise removal is performed using the morphological opening operation.

B. Concavity detection

After preprocessing, every cluster in the image is treated separately. Figure 2 illustrates cluster segmentation steps for a cluster. A concavity is the point on the cluster edge where two individual nuclei overlap and is therefore the point where a cluster should be segmented. Concavity detection can be carried out using the cross product of adjacent tangential vectors while moving along the edge in one direction [7]. In this method, we divide the cluster edge into piecewise segments and determine the tangential vectors for every segment using endpoints, given by

$$\mathbf{a} = [p_{2x} - p_{1x},\ p_{2y} - p_{1y}, 0]\ ;\ \mathbf{b} = [p_{3x} - p_{2x},\ p_{3y} - p_{2y}, 0] \quad (2)$$

where **a** and **b** are two adjacent tangential vectors defined by three adjacent points on the cluster edge, p_1, p_2, and p_3. In our case, when the *z*-component is zero for the **a** and **b** vectors, the cross-product extends in the *z*-direction and sin(*θ*) can be given by

$$\sin(\theta) = \frac{1}{|\mathbf{a}||\mathbf{b}|}[\mathbf{a}_x\mathbf{b}_y - \mathbf{a}_y\mathbf{b}_x] \quad (3)$$

The value of sin(*θ*) is in the positive *z*-direction for concave portions of the edge and in the negative *z*-direction for convex portions as shown in Figure 2(a). Figure 2(b) illustrates the variation of sin(*θ*) along the cluster edge. We can find the concavity by detecting zero-crossings of sin(*θ*) and finding maxima between two crossings. As compared to previous work [7], this methodology doesn't need a fixed threshold.

C. Straight line segmentation

Straight line segmentation is the first step in the segmentation of clusters. First, the user interactively selects possible single nuclei. From these, the approximate nuclei size, *A,* is calculated. Straight line (SL) segmentation of clusters is an iterative process. At each step, there is a check on the size of resulting nuclei compared to *A*, given by

$$\textit{Resulting nuclei area} > \textit{threshold} \times A. \quad (4)$$

The threshold is decided depending on single nuclei-size variation in the tissue image. We obtained good results for different types of images using a value of 0.3. For SL segmentation, we calculate the distance between all concavities for a cluster and connect the concavities starting with the ones closest to each other. The concavities are connected only if the larger portion of the connecting line segment lies inside the cluster. Figure 2(c) depicts the SL segmented regions of the cluster in Figure 2(a). SL segmentation may lead to over segmentation of the cluster, but ellipse fitting overcomes this issue.

D. Ellipse fitting

We adopt an elliptical model for nuclei [6]. We have used a direct ellipse fitting method proposed by Fitzgibbon et al. [11] due to its accuracy and simplicity of implementation. We propose an improved methodology of using direct ellipse fitting on regions of SL segmented cluster as

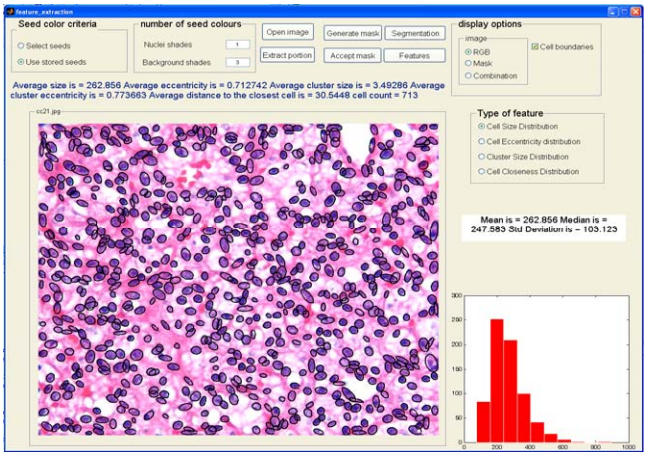

Figure 4: Snapshot of GUI after cluster segmentation

compared to previous work [7].

Firstly, we sort the regions of a cluster from SL segmentation in decreasing order of precedence depending on the portion of the cluster edge it includes. We start ellipse fitting with the region with highest precedence. The data provided to the ellipse fitting algorithm is the portion of SL segmented region that is a part of the cluster edge. If this data is insufficient or lies on a straight line, then the complete edge of the SL segmented region is provided.

Ellipse fitting is an iterative process. At every step we check the overlap between the present ellipse and previously fitted ellipses. To avoid over segmentation, only ellipses with a larger portion of non-overlapping regions are selected. To reduce the number of missed detections, once the ellipse fitting is completed for all clusters, we check if there is any portion of the mask which was not considered a nucleus but is of sufficient size to be one. The boundaries of such regions are provided to the ellipse fitting algorithm. Figure 2(d) shows the result of ellipse fitting. Figure 3(c) & 3(d) shows final segmentation results of the tissue sample in figure 3(a).

IV. Informative Feature Extraction

After ellipse fitting is completed, every nucleus has been modeled as an ellipse using basic parameters such as major axis a, minor axis b, center (x, y) and orientation α. With the help of these parameters, we can extract multiple informative features of the tissue image. We illustrate four such features in our result. The first three depend on individual nuclei parameters and the last depends on the aggregation pattern of the nuclei.

A. Nuclei size distribution

Nucleus size can be approximated by calculating the area of the elliptical nucleus, given by

$$Cell\ Area = \pi \times a \times b \tag{5}$$

Nuclei size distribution for the complete tissue image is estimated by calculating areas for all nuclei.

B. Nuclei shape distribution

For nuclei with a purely elliptical model, one of the important measures for shape is eccentricity, given by

$$Cell\ eccentricity = \sqrt{1 - \frac{b^2}{a^2}} \tag{6}$$

If the eccentricity of the nucleus is close to zero, then the nucleus is more circular.

C. Nuclei closeness distribution

The average distance between every nucleus and the nuclei in its neighborhood estimates the overall closeness of nuclei in the tissue image. We calculate the average distance of every nucleus to its five closest nuclei for closeness measurement, given by

$$D = \frac{1}{5} \times \sum_{i=1}^{5} \sqrt{(x_i - x)^2 + (y_i - y)^2} \tag{7}$$

where (x_i, y_i) are the five closest nuclei centers and (x, y) is the center of the five nuclei.

D. Cluster size distribution

In certain cases, even when the average closeness measure of nuclei is similar, the images appear to be different based on the size of the clusters. We tried to capture this feature of the images by providing a measure of cluster size in terms of the number of nuclei contained in each cluster.

V. Graphical User Interface

A simple GUI was developed in Matlab for this application. The GUI works in a sequential order and we provide the user with step by step instructions and enable relevant push-buttons at each step. After selecting an image, the user's involvement is required at only three steps: 1) selecting the number of seed colors for background shades and nuclear shades and selecting the seeds themselves, 2) adjusting the nuclei mask by varying the means of the clusters, and 3) clicking on samples of single nuclei for approximating nuclei size. After cluster segmentation, the user can view distribution histograms of different features, including their mean, median, and standard deviation. Also, the user can view segmented nuclei boundaries on the desired image (RGB tissue image, binary mask and combination options are available). Figure 4 shows a snapshot of the GUI after cluster segmentation.

VI. Results

In order to show the importance of the features extracted by our methodology, we compared the features of four tissue images. Out of these four images, two are renal cell carcinoma (RCC) tissue images and other two are head & neck (H&N) cancer images. In addition, to evaluate repeatability of the technique, we manually selected four sub-portions of every tissue image and extracted features for each sub-portion. Figure 5(a) - (d) shows one sub portion from each image with their respective time delay and cell count. The computation time for cluster segmentation is dependent on the number of cells in the image. The average computation time was less than 2 minutes for any image in a dataset of 58 1.9 megapixel RCC images.

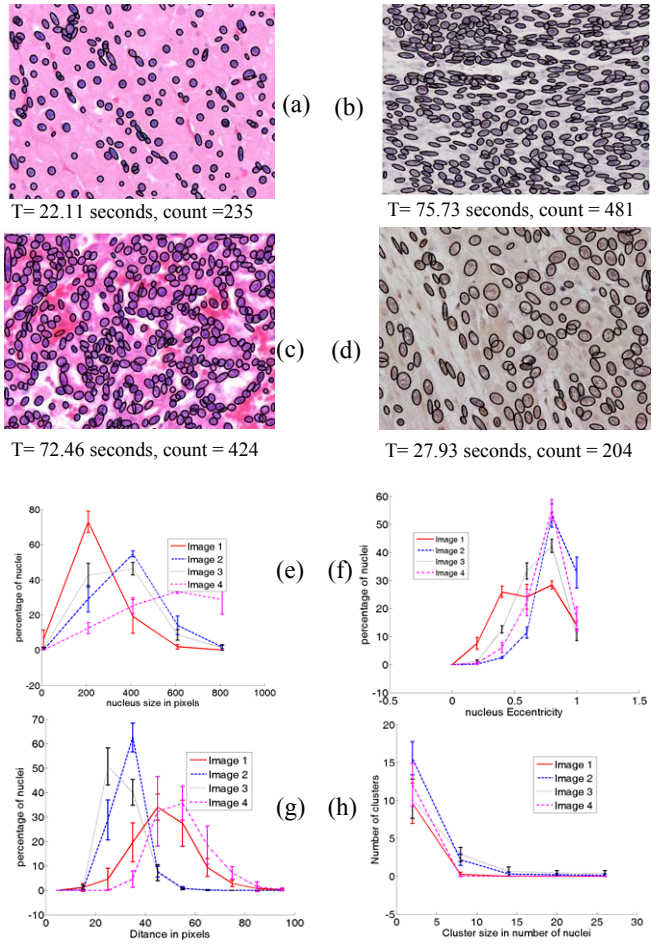

Figure 5: Comparison of features of four tissue images. (a)-(d) image 1-4 in comparison (e)-(h) with individual cells marked, (e) nuclei size distribution, (f) nuclei shape distribution, (g) nuclei closeness distribution, (h) cluster size distribution

We calculated the average distribution for every tissue image from its sub-portions, and then we plotted comparison graphs between the four tissue images for all four features. The first graph in figure 5(e) compares the nuclei size distribution in four tissue images. The average nuclei size is smallest in the first image and largest in the fourth image, therefore the size distribution for the first image has its maxima at a very lower value as compared to the fourth image. The second graph in figure 5(f) compares nuclei shape distributions measured as the eccentricity of nuclei. Image 2 and image 4 have a higher percentage of elliptical nuclei as compared to circular nuclei. As such, their nuclei shape distribution maxima is closer to one. Moreover, image 1 has mostly circular nuclei; as such, its distribution has a higher percentage of nuclei for lower eccentricities compared to other images. The third graph in figure 5(g) compares the nuclei closeness distribution of four images. Nuclei are closely distributed in image 2 and image 3 compared to image 1 and image 4, therefore the maxima of their distribution is lower on the x-axis. The fourth graph in figure 5(h) compares the aggregation patterns of nuclei in four images. Clustering of nuclei is greater in image 2 and image 3. Therefore, the distributions of cluster size for image 2 and image 3 are higher at larger size clusters separating them from the other two images. Also, only image 3 has long clusters, so its distribution is present for higher values on the *x*-axis. The four tissue images can be differentiated using these comparison graphs. Image 3 can be differentiated from all other images using cluster size distribution. Image 1 can be differentiated using nuclei size distribution. Image 2 and image 4 can be differentiated from all other images using nuclei eccentricity distribution and from each other using nuclei closeness distribution.

VII. Conclusion

We have considered a set of four images in our results and provided users with sample images as well as feature comparison graphs. As an extension of this work, we plan to extract features from a larger image dataset and classify the images using these features. We have shown that, with good nuclei cluster segmentation and elliptical modeling of nuclei, we can generate informative image features. This work can be extremely helpful in image classification and image grading of pathological images in clinical settings.

Acknowledgment

We thank Dr. Mitch Parry and Richard Moffitt for their valuable comments and suggestions.

References

[1] C. François et al, “Improving accuracy in the grading of renal cell carcinoma by combining the quantitative description of chromatin pattern with the quantitative determination of cell kinetic parameters,” Cytometry, vol. 42, pp. 18-26, 2000.

[2] V. Kirillov et al, “Thyroid carcinoma diagnosis based on a set of karyometric parameters of follicular cells,” Cancer, vol. 92, pp. 1818-1827, 2001.

[3] J. Gil, H. Wu, and B. Wang, “Image analysis and morphometry in the diagnosis of breast cancer,” in Microscopy Research and Technique, vol. 59, pp. 109-118, 2002.

[4] E.Glory et al, “Automated image-based screening of cell cultures for cell therapy”, Biomedical Imaging: Nano to Macro., IEEE, pp. 259-262 , 2006

[5] Nilsson, B., Heyden, A., “Segmentation of dense leukocyte clusters” , CVPR, IEEE, pp. 221-227, 2001

[6] X. Bai et al, “Touching Cells Splitting by Using Concave Points and Ellipse Fitting”, DICTA, IEEE, pp.271-278, Dec. 2008 Nilsson, B., Heyden, A., “Segmentation of dense leukocyte clusters” , CVPR, IEEE, pp. 221-227, 2001

[7] S. Kothari et al, “Automated cell counting and cluster segmentation using concavity detection and ellipse fitting techniques,”, IEEE International Symposium on Biomedical Imaging (ISBI’09), accepted for publication.

[8] A.R. Weeks, and G.E. Hague, “Color Segmentation in the HSI Color Space Using the k-means Algorithm," Proc. of the SPIE - Nonlinear Image Processing VIII, pp. 143-154, 1997

[9] Q. Chaudry et al, “Improving renal cell carcinoma classification by automatic region of interest selection”, BioInformatics and BioEngineering, IEEE, pp. 1-6, October, 2008

[10] Soille, P., Morphological Image Analysis: Principles and Applications, Springer-Verlag, pp. 173-174, 1999

[11] A. Fitzgibbon et al, “Direct least square fitting of ellipses”, Transactions on Pattern Analysis and Machine Intelligence, IEEE, Volume 21, pp. 476 – 480, May 1999