AUTOMATIC BATCH-INVARIANT COLOR SEGMENTATION OF HISTOLOGICAL CANCER IMAGES

Sonal Kothari[1], John H. Phan[2], Richard A. Moffitt[2], Todd H. Stokes[2], Shelby E. Hassberger[2], Qaiser Chaudry[1], Andrew N. Young[3,4], May D. Wang[2]

[1]Electrical and Computer Engineering, Georgia Institute of Technology
[2]Biomedical Engineering, Georgia Institute of Technology and Emory University
[3]Pathology and Laboratory Medicine, Emory University, [4]Grady Health System, Atlanta, GA
([*]maywang@bme.gatech.edu)

ABSTRACT

We propose an automatic color segmentation system that (1) incorporates domain knowledge to guide histological image segmentation and (2) normalizes images to reduce sensitivity to batch effects. Color segmentation is an important, yet difficult, component of image-based diagnostic systems. User-interactive guidance by domain experts—i.e., pathologists—often leads to the best color segmentation or "ground truth" regardless of stain color variations in different batches. However, such guidance limits the objectivity, reproducibility and speed of diagnostic systems. Our system uses knowledge from pre-segmented reference images to normalize and classify pixels in patient images. The system then refines the segmentation by re-classifying pixels in the original color space. We test our system on four batches of H&E stained images and, in comparison to a system with no normalization (39% average accuracy), we obtain an average segmentation accuracy of 85%.

Index Terms— color segmentation, supervised learning, normalization, histological images.

1. INTRODUCTION

Color-enhanced, or stained, cellular structures in histological images enable clinicians to identify morphological markers of a disease, and to proceed with therapy accordingly. However, because of variations in specimen preparation, staining, and imaging, resulting images may exhibit very different colors (Figure 1). Under such conditions, computer-aided diagnostic systems [1-4] that segment these structures based on their color often fail. One way to account for the observed difference in colors among images, i.e. 'batch effect', is to develop an interactive system that allows users to lend their domain knowledge to guide the segmentation process [1, 2]. However, user-interaction lowers the overall objectivity, reproducibility and speed of such systems. Among automatic segmentation methods, supervised learning techniques have been reported to be more accurate than unsupervised learning methods [5-7]. We find that these previous techniques are vulnerable to batch effect, and that they tend to perform well only for data from the batch on which they are trained (Table 1). Therefore, we propose a system for automatic color segmentation of histological images which is designed to be resistant to batch effect (Figure 2). Our system incorporates knowledge from pre-segmented reference images to normalize (Figure 2, Step 1) and segment (Figure 2, Step 2) new patient images. Also, in order to make our system robust to the choice of reference image (j), we segment new images (k) with multiple reference images and combine labels, $\mathbf{L}_{0,j}^{k}$, using a voting scheme. Voting produces preliminary segmentation labels, $\mathbf{L}_{1}^{k}$, which we then use to reclassify (Figure 2, Step 3) test image pixels in their original color space and produce final segmentation labels, $\mathbf{L}_{2}^{k}$. The proposed system provides an automatic color segmentation of histopathological specimens that is resistant to batch effects. We achieve this by incorporating knowledge from domain experts into a novel color normalization scheme.

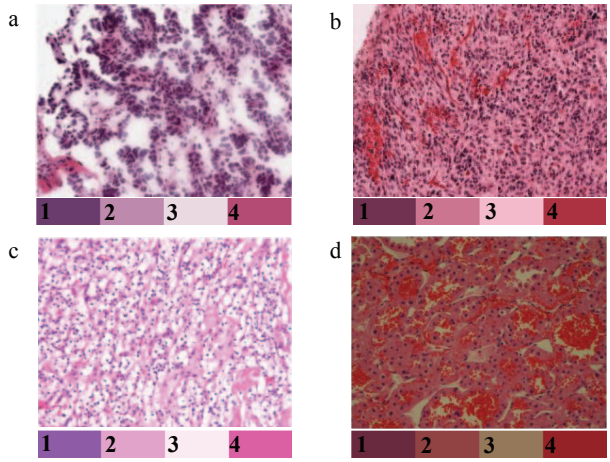

Figure 1. Sample images from four datasets: a. Ovarian (Ov) , b. Glioblastoma (Gbm), c. Renal tumor (RCC1), d. Renal tumor (RCC2). Color palette illustrates cluster means (in ground truth) of four color classes- 1) blue-purple (basophilic), 2) pink (eosinophilic), 3) white (no stain) and 4) red (red blood cells).

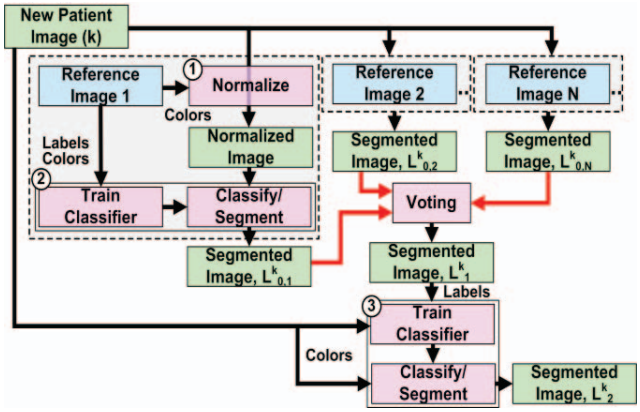

Figure 2. System flow diagram with three main steps: 1) normalize, 2) segment normalized image, 3) re-classify pixels in the original color space.

2. METHODS

2.1. Datasets

We analyze photomicrographs of histological specimens stained with hematoxylin and eosin (H&E). Basophilic structures containing nucleic acids—ribosomes and nuclei—tend to stain blue-purple; eosinophilic intra- and extra-cellular proteins tend to stain the cytoplasm pink; empty spaces—the lumen of glands—do not stain and tend to be white; and red blood cells appear intensely red. Thus, H&E staining produces four distinguishable clusters of colors in the image: blue-purple, white, pink and red. The color palettes in Figure 1 illustrate the mean color for each of the four color clusters in the ground truth segmentation. We consider four datasets (Figure 1): two renal tumor (RCC1 and RCC2 with 55 and 47 images, respectively), one glioblastoma (Gbm, 52 images), and one ovarian (Ov, 50 images). RCC1 and RCC2 were obtained at Emory University in separate experimental setups. Ov and Gbm images were obtained from the NIH's public Cancer Genome Atlas (TCGA) repository. We use 1024x1024-pixel cropped portions of the original slide images. All datasets have varying grades and subtypes of cancer, leading to changes in the morphology and thus the prevalence of the four color classes in the images. Therefore, besides a color-based batch effect, the data sets also have a prevalence-based batch effect.

2.2. Ground Truth Preparation

To establish the ground truth labeling for each image, we developed an interface to help users label pixels semi-automatically. In this system, the user first selects four example pixels from an image to represent the four stain colors. Each remaining pixel is then classified into one of four groups based on minimum Euclidean distance in the RGB color space to the four example pixels. The user may then fine-tune this segmentation by using a slider-bar to adjust weights given to these distances. This effectively adjusts the dominance of each cluster in the segmentation until the user is satisfied. We use the final labels from this segmentation to prepare reference images and to assess performance.

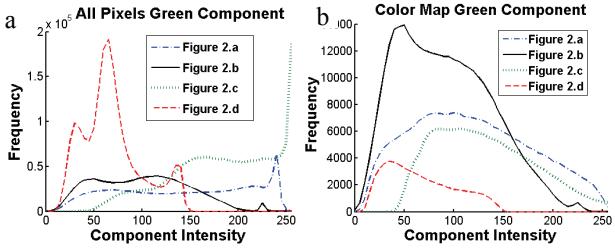

Figure 3. Distribution of green component intensities for (a) *all pixels* and (b) *color map* of images in Figure 1. Compared to *color map*, *all pixels* contain peaks which vary with changes in morphology and class prevalence.

2.3. Image Normalization

We begin segmenting sample images by first normalizing the sample image's colors to the reference image's colors. Many color normalization techniques have been proposed [8, 9], including histogram or quantile normalization in which the distributions of the three color channels are normalized separately. Here, we mathematically describe quantile normalization of *all pixels* in an image. An image k contains N_k pixels where each pixel n is represented as a triplet, $I^{k,n} = [R^{k,n}, G^{k,n}, B^{k,n}]$. $R^{k,n}, G^{k,n}$, $B^{k,n}$ are color channel intensity values. We define a rank function $\mathrm{f}_C^k \in \Re^{N_k \times N_k}$ that maps the color channel intensity, $C \in [R,G,B]$, from image k to a rank that ranges from 0 to N_k-1. Using the green channel as an example, $\mathrm{f}_G^k(\mathbf{G}^k) = \mathbf{r}_G^k$, where $\mathbf{G}^k, \mathbf{r}_G^k \in \Re^{N_k}$ are vectors of the green component intensity and rank for the k^{th} image, respectively. If $G^{k,n} \in [0,255]$ and $r_G^{k,n} \in [0, N_k - 1]$ are green component intensity and rank for the n^{th} pixel in the k^{th} image, then for any two pixels n_1 and n_2, $r_G^{k,n_1} \le r_G^{k,n_2}$ *iff* $G^{k,n_1} \le G^{k,n_2}$. The normalized green channel intensity of the n^{th} pixel of the k^{th} sample image to the j^{th} reference image can be computed with $\tilde{G}_j^{k,n} = \mathrm{h}_G^j\left\lfloor \frac{r_G^{k,n}}{N_k} \times N_j + \frac{1}{2} \right\rfloor$, where $\mathrm{h}_G^j\left(r_G^{j,n}\right) = G^{j,n}$ is the inverse of the j^{th} image's green rank function $\mathrm{f}_G^j(\mathbf{G}^j) = \mathbf{r}_G^j$.

We propose an alternative to simple quantile normalization where we use the *color map* of the image instead of *all pixels* in the image. The *color map* is obtained by extracting the unique colors in the image. Therefore, compared to *all pixels*, the *color map* does not include the frequency of any colors. Mathematically, quantile normalization of *color map* elements is similar to that of *all pixels* except that the image is represented by a list of unique color triplets, $U^{k,m} = [R^{k,m}, G^{k,m}, B^{k,m}]$, where $m \in [0, M_k - 1]$ and M_k is the number of unique colors in the image. Because of variations in morphology from image to

Table 1: Segmentation accuracy compared to *ground truth*.

Train	Test	No norm.		*All pixels*		*Color map*	
		L_1	L_2	L_1	L_2	L_1	L_2
RCC1	RCC2	0.17	0.08	0.83	0.82	0.79	0.82
	Ov	0.32	0.43	0.79	0.83	0.77	0.81
	Gbm	0.22	0.54	0.82	0.85	0.80	0.82
RCC2	RCC1	0.37	0.56	0.85	0.88	0.79	0.87
	Ov	0.32	0.40	0.82	0.85	0.86	0.87
	Gbm	0.23	0.46	0.80	0.84	0.81	0.84
Ov	RCC1	0.13	0.13	0.73	0.78	0.82	0.87
	RCC2	0.16	0.16	0.72	0.74	0.85	0.84
	Gbm	0.77	0.82	0.76	0.80	0.85	0.85
Gbm	RCC1	0.13	0.13	0.82	0.84	0.85	0.87
	RCC2	0.16	0.16	0.78	0.80	0.84	0.83
	Ov	0.84	0.84	0.78	0.83	0.87	0.87
Overall		**0.32**	**0.39**	**0.79**	**0.82**	**0.82**	**0.85**

* p-value for t-tests between: 1) L_2 *all pixels* and L_2 *color map* is—0.044, 2) L_1 and L_2 *color map* is—0.010.

image, color and class frequencies vary. Figure 3 illustrates the distribution of green component intensity for *all pixels* and for *color map* elements of the four images in Figure 1. While the distributions of *all pixels* contain peaks which vary with changes in morphology and class prevalence, the distributions of *color map* shows less change between images. Therefore, normalizing the *all pixels* distributions rather than the *color map* distributions tends to distort colors in the normalized image. Once colors have been normalized, pixels are then classified by color.

2.4. Normalized Image Segmentation

Pixel classification is performed in the color space of a reference image. Using a four-class linear discriminant classifier (LDA), we train using colors and labels obtained from ground truth segmentation of the reference image and classify pixels from the sample images based on normalized color. Let $\mathbf{L}^j \in \Re^{N_j}$ and $\mathbf{I}^j \in \Re^{N_j \times 3}$, where each element is $I^{j,n} = [R^{j,n}, G^{j,n}, B^{j,n}]$, be defined as the user-interactive segmentation labels and color values of pixels in image *j*, respectively. Let $\tilde{\mathbf{I}}_j^k \in \Re^{N_k \times 3}$ be defined as image *k* normalized to image *j* where each element is given by $\tilde{I}_j^{k,n} = [\tilde{R}_j^{k,n}, \tilde{G}_j^{k,n}, \tilde{B}_j^{k,n}]$. For convenience, we define the function $\mathbf{L}' = \mathrm{LDA}(\mathbf{I}^j, \mathbf{L}^j, \mathbf{I}^k)$, where $\mathbf{L}'$ contains segmentation labels for image $\mathbf{I}^k$ using an LDA classifier trained with pixel colors in $\mathbf{I}^j$ and labels in $\mathbf{L}^j$. $\mathbf{I}^k$ may also be a normalized image. Thus, to obtain the segmented image labels, $\mathbf{L}_0$ (Figure 2), we use $\mathbf{L}_{\mathbf{0,j}}^{\mathbf{k}} = \mathrm{LDA}(\mathbf{I}^{\mathbf{j}}, \mathbf{L}^{\mathbf{j}}, \tilde{\mathbf{I}}_{\mathbf{j}}^{\mathbf{k}})$.

The accuracy of segmentation depends on the choice of reference image. Due to biological variation in patients, reference image colors within a single dataset tend to vary, affecting color normalization and segmentation. Therefore, in order to select optimal reference images, we perform cross validation within each dataset batch. We use each image in the batch as a reference to normalize and segment

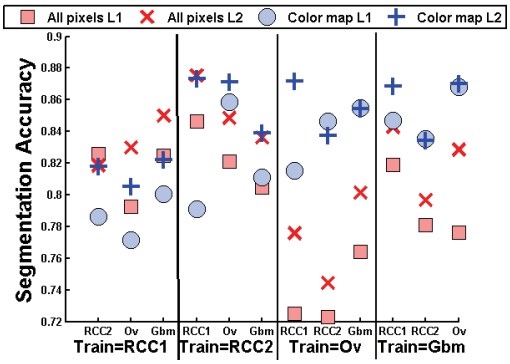

Figure 4. Comparison of segmentation accuracy of *all pixels* L_1, *all pixels* L_2, *color map* L_1, *color map* L_2.

all remaining images in the batch. The performance value of each reference image is the average segmentation accuracy of all remaining images. We select the top 10 performing references from each batch. This methodology selects reference images based on their ability to normalize and accurately segment other images in the batch. We did not observe significant changes in performance when selecting more or less than 10 reference images.

In order to avoid the choice of a single canonical reference image, we develop a system that allows the use of multiple reference images. In our system, a sample image is normalized and segmented 10 times, using a different reference image each time. For each pixel in the sample image, we compute the final segmentation label by voting from multiple references. The label most frequently assigned to a pixel is chosen as its preliminary label (block $\mathbf{L}_1^k$ in Figure 2) before segmentation refinement.

2.5. Segmentation Refinement

The preliminary labels obtained by classification and voting ($\mathbf{L}_1^k$, Figure 2) are good approximations of the ground truth labels, but we further refine this segmentation using the LDA classifier: $\mathbf{L}_2^k = \mathrm{LDA}(\mathbf{I}^k, \mathbf{L}_1^k, \mathbf{I}^k)$. This step trains the LDA classifier using colors from the original sample image *k* and using labels from voting. The trained classifier is then used to re-classify all pixels in image *k*. Intuitively, this is a post-processing step that ensures that the color groupings are separable in the original sample image color space, and that any color distortion introduced by normalization is removed.

3. RESULTS AND DISCUSSION

Table 1 lists the segmentation results from our system using two types of normalization (*all pixels* or *color map*) and compares them to our system with no normalization, i.e. $\mathbf{L}_{0,j}^k = \mathrm{LDA}(\mathbf{I}^j, \mathbf{L}^j, \mathbf{I}^k)$. We report accuracy of both $\mathbf{L}_1$ and $\mathbf{L}_2$ labels, to illustrate how re-classification of pixels in the original color space of the sample image improves segmentation accuracy. The overall performance, at 85%

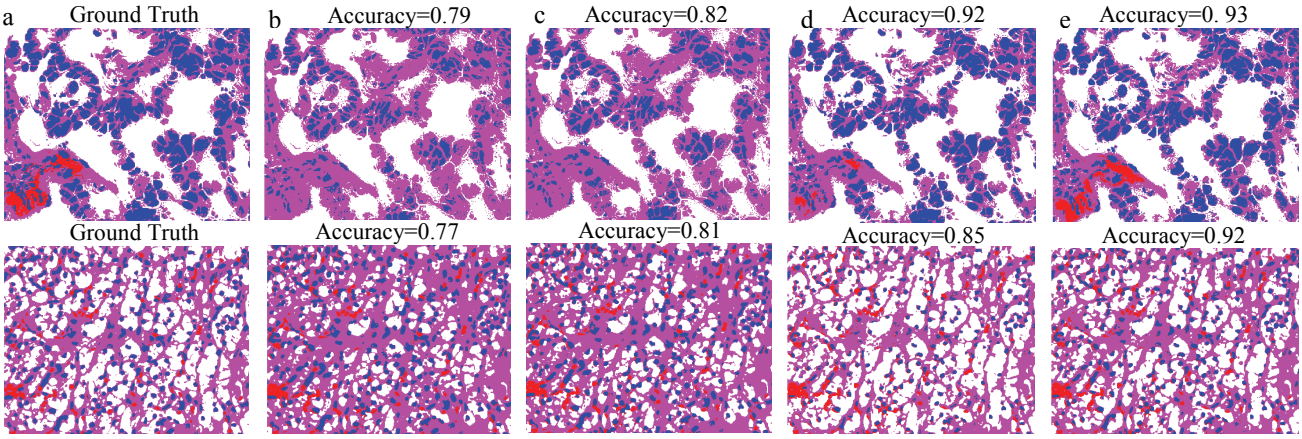

Figure 5. Segmentation of the images in Figure 1.a (top) and Figure 1.c (bottom). A magnified lower left portion of the image is shown; however, accuracy is reported for the full image. a) ground truth, b) *All pixels* L1, c) *All pixels* L2, d) *Color map* L1, e) *Color map* L2.

accuracy, is best for a system that uses *color map* normalization and re-classification (*color map* L2). Figure 4 compares segmentation results with *color map* and *all pixels* normalization. Re-classification (Figure 4, + and x) significantly improves the segmentation performance. *Color map* normalization performs better than *all pixels* normalization except for four cases involving the RCC1 batch, possibly due to chromatic aberration, resulting in color map histogram distortion. However, in *all pixels* normalization, due to the low frequency of chromatic aberration colors, distortion is less severe. Figure 5 illustrates pseudo colored segmentation results for images in Figure 1.a and Figure 1.c. Figure 1.a is an Ov batch image and is segmented on a system trained by reference images from the RCC2 batch. Figure 1.c is an RCC1 batch image and is segmented on a system trained by reference images from the Gbm batch. Again, the re-classification step enhances the segmentation results and *color map* normalization retains the morphology of the test image. For instance, in the second row of Figure 5, *all pixels* normalization alters the morphology of the test image, over-segments the pink mask, and under-segments the white mask. Similarly in the top row of Figure 5, the pink mask is over-segmented while the other three masks are under-segmented.

Automatic computer-aided cancer diagnostic systems for histological images are necessary to improve objectivity, reproducibility and speed of diagnosis. Many systems [1-4] require a color-based segmentation, which is often sensitive to batch effects. In this paper, we have presented an automatic color segmentation system that uses an expert's initial domain knowledge to normalize image colors prior to segmentation in order to reduce the effect of variance between batches of images. The high accuracy of these segmentation masks, relative to expert domain knowledge, will aid in increasing the overall performance and reproducibility of diagnostic systems that depend on color segmentation. We have tested our system using four very different batches of H&E histological cancer images and achieved high segmentation accuracy (85%). We expect that our system can be extended to other staining protocols, e.g. Papanicolaou stain, and is not restricted to segmentation problems with four stain colors.

4. ACKNOWLEDGMENT

This research has been supported by grants from NIH (Bioengineering Research Partnership R01CA108468, P20GM072069, and CCNE U54CA119338), Georgia Cancer Coalition, Hewlett Packard, and Microsoft Research.

5. REFERENCES

[1] S. Kothari, *et al.*, "Extraction of informative cell features by segmentation of densely clustered tissue images," in *Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE*, 2009, pp. 6706-6709.

[2] Q. Chaudry, *et al.*, "Automated Renal Cell Carcinoma Subtype Classification Using Morphological, Textural and Wavelets Based Features," *Journal of Signal Processing Systems,* vol. 55, pp. 15-23, 2009.

[3] A. Basavanhally, *et al.*, "Computerized image-based detection and grading of lymphocytic infiltration in her2+ breast cancer histopathology," *Biomedical Engineering, IEEE Transactions on,* vol. 57, pp. 642-653, 2010.

[4] A. Tabesh, *et al.*, "Multifeature prostate cancer diagnosis and Gleason grading of histological images," *IEEE Transactions on Medical Imaging,* vol. 26, pp. 1366-1378, 2007.

[5] C. Meurie, *et al.*, "A comparison of supervised pixels-based color image segmentation methods. application in cancerology," *WSEAS Transactions on Computers,* vol. 2, pp. 739-44, 2003.

[6] K. Mao, *et al.*, "Supervised learning-based cell image segmentation for P53 immunohistochemistry," *Biomedical Engineering, IEEE Transactions on,* vol. 53, pp. 1153-1163, 2006.

[7] P. Ranefalla, *et al.*, "A new method for segmentation of colour images applied to immunohistochemically stained cell nuclei," *Analytical Cellular Pathology,* vol. 15, pp. 145-156, 1997.

[8] D. Magee, *et al.*, "Colour Normalisation in Digital Histopathology Images," in *Proc. Optical Tissue Image analysis in Microscopy, Histopathology and Endoscopy (MICCAI Workshop)*, 2009, pp. 100-111.

[9] M. Macenko, *et al.*, "A method for normalizing histology slides for quantitative analysis," 2009, pp. 1107-1110.