Published in final edited form as:
Proc SPIE Int Soc Opt Eng. 2017 February ; 10140: . doi:10.1117/12.2254220.

Evaluation of nucleus segmentation in digital pathology images through large scale image synthesis

Naiyun Zhou[d], **Xiaxia Yu**[a], **Tianhao Zhao**[a], **Si Wen**[c], **Fusheng Wang**[a,b], **Wei Zhu**[c], **Tahsin Kurc**[a,b], **Allen Tannenbaum**[b,c], **Joel Saltz**[a,b], and **Yi Gao**[a,c]

[a]Department of Biomedical Informatics, Stony Brook University

[b]Department of Computer Science, Stony Brook University

[c]Department of Applied Mathematics and Statistics, Stony Brook University

[d]Department of Biomedical Engineering, Stony Brook University

Abstract

Digital histopathology images with more than 1 Gigapixel are drawing more and more attention in clinical, biomedical research, and computer vision fields. Among the multiple observable features spanning multiple scales in the pathology images, the nuclear morphology is one of the central criteria for diagnosis and grading. As a result it is also the mostly studied target in image computing. Large amount of research papers have devoted to the problem of extracting nuclei from digital pathology images, which is the foundation of any further correlation study. However, the validation and evaluation of nucleus extraction have yet been formulated rigorously and systematically. Some researches report a human verified segmentation with thousands of nuclei, whereas a single whole slide image may contain up to million. The main obstacle lies in the difficulty of obtaining such a large number of validated nuclei, which is essentially an impossible task for pathologist. We propose a systematic validation and evaluation approach based on large scale image synthesis. This could facilitate a more quantitatively validated study for current and future histopathology image analysis field.

Keywords

histopathology; nucleus extraction; segmentation evaluation; image synthesis

1. INTRODUCTION

Digital histopathology image computing is becoming a more and more active field due to its superb spatial resolution and the availability of large data sets. At a spatial scale between genomics and radiology, histopathology images have shown promising possibility of revealing detailed image based features that can be linked with genetics in a straightforward manner, as well as the epigenetical and environmental factors, which are not present in the DNA sequences.

Further author information: Send correspondence to Dr. Yi Gao. Yi.Gao@stonybrookmedicine.edu.

Among the vast amount of image features in the digital pathology images, the nuclei lie at a central position. Their morphologies are the main criteria for pathologist making diagnostic and grading decisions. Image computing often starts with nucleus analysis, the accurate extraction of each of every nuclei being the pre-requisite of such analysis.[1]

There exist large volume of literature discussing nucleus segmentation.[2–12] The quantitative validation and evaluation of the nucleus segmentation algorithm is one of the most challenging issue in all of the publications. Indeed, it is difficult, if not impossible, to ask a skilled pathologist to manually contour hundreds and thousands of nuclei to be compared with algorithm output. Even so, one whole slide image may contain up to a million nuclei. Therefore, instead of relying on human annotation, we need a systematic approach to computationally generate large enough data set to test the accuracy and robustness of the algorithm, which is the main topic of this paper.

2. METHODS

2.1 Image Synthesis and Quantitative Evaluation Framework

To set the stage up, the algorithm starts with a set of images $I_i : \Omega \subset \mathbb{R}^2 \rightarrow$ C; i = 1, 2, …, N, where C denotes the RGB color space. We have expert validated segmentation of the nuclei in I_i’s, denoted as label images $J_i : \Omega \rightarrow \{0\} \cup \mathbb{Z}^+$. In the label images, 0 denotes background and different integer values indicate different nuclei. Denote the bounding box of the j-th nucleus in I_i as $B_i^j : = (x_i^j, y_i^j, u_i^j, v_i^j); j = 1, 2, \ldots, N_i$, where (x, y)($resp.$$(u, v)$) denotes the top-left (resp. bottom-right) coordinates for the bounding box, and N_i is the number of nuclei in I_i. With slight abuse of notation, the image domain of the bounding box is also denoted as B_i^j. Furthermore, the restriction of the image and the label iamge, i.e., the patch around the nucleus, are denoted as $I_i^j : = I_{i | B_i^j}$ and $J_i^j : = J_{i | B_i^j}$, respectively. Finally, the non-nuclear region in the patch is removed by defining $I_i^j : = I_i^j * J_i^j$. To simplify the notation, we use single index and denote $P_1 = I_i^1, P_2 = I_i^2$ and $Q_1 = J_i^1, Q_2 = J_i^2$ etc.

2.1.1 Nucleus Synthesis—In this step, we synthesize new nuclei based on those in the training images. To proceed, a random nucleus P is picked from all P_i’s. Together with its mask Q, they will be used to generate a new nucleus. In order to generate nuclei with different orientation and thus evaluate the *orientation invatiance* of the segmentation algorithm, the patch P and Q are rotated with a random degree $\theta \in [0, 2\pi)$. The size of nucleus is also a key factor being considered in clinical practice. To simulate the *size variation*, the sizes of the training nuclei are computed and the size distribution, p_S, is estimated through a kernel density estimation process.[13] Then, a sample is drawn from p_S and the new nucleus is scaled to the size.[14]

After that, a nonlinear deformation is applied to the new nucleus. At the same time, the corresponding label map is also deformed. As a result, we still have the exact segmentation of the nucleus. To this end, one possible approach is to generate a random deformation field and smooth it to certain degree. Such a random deformation may be feasible from a computational point of view. Yet, the random deformation may generate some unrealistic shape. Because of this, we propose a relative conservative approach so that the resulting deformed nucleus is within the space spanned by the training nuclei. Specifically, for every training nuclei, without loss of generality, we assume their sizes (areas) have been normalized. Then, a Procrustes alignment is performed so that all major axes are aligned north-south. After that, each nucleus is mapped to the unit disk through the optimal mass transport (OMT).[15,16] Formally, a set of points $G := \{x_i \in \mathbb{R}^2$: i = 1, 2, …, m} is sampled uniformly from a unit disk, and a same number of points $H := \{y_i\, \mathbb{R}^2$: i = 1, 2, …, m} are sampled from the nucleus to be registered to the disk. Each point is considered to have a Dirac metric. Then, an optimal matching between the two sets of points is constructed. To that end, we denote the correspondence between X and Y by a matrix $A \in \{0, 1\}^{m \times n}$, where $A_{i,j} = 1(0)$ indicates x_i is corresponding (not corresponding, resp.) with y_j. Denoting the pair-wise distance matrix $C \in \mathbb{R}^{\mathrm{m} \times \mathrm{m}}$ as $C_{i,j} = \left\|x_i - y_j\right\|_2$, where $\left\| \cdot \right\|_2$ is the L_2 norm, we find the correspondence between the two sets of points by solving such an assignment problem:

$$\begin{aligned} & A = \min_{\tilde{A} \in \mathbb{R}^{\mathrm{m} \times \mathrm{m}}} \left\|C \circ \tilde{A}\right\|_F \qquad (1) \\ & s.t. \sum_j \tilde{A}_{i,j} = 1 \quad \forall i \in \{1, \ldots, m\} \\ & \sum_j \tilde{A}_{i,j} = 1 \quad \forall i \in \{1, \ldots, m\} \\ & \tilde{A}_{i,j} \geq 0 \quad \forall i, j \in \{1, \ldots, m\} \end{aligned}$$

where is the Hadamard product of the two matrices and $\left\| \cdot \right\|_F$ is the matrix Frobenius norm.

Moreover, it is noted that the optimization variable Ã is not restricted to be a binary matrix. Otherwise the optimization becomes an NP-hard combinatorial problem. On the other hand, due to the fact that the constraint matrix of (1) is totally unimodular, the resulting optimal A is a binary matrix.[17] This optimization problem can be shown to be convex, and it can be effectively solved by using, for example, interior point method.[18] The resulting matrix A will give a one-to-one correspondence between G and H. The deformation field D_i is therefore constructed as the displacement vector field among the corresponding points.

In a high dimensional deformation space, the D_i's most likely reside on a manifold rather than in a linear space.[19,20] In order to generate arbitrarily deformation with similar fashion to that has been observed in the training data, one can interpolate the deformation fields on the manifold. However, due to the high dimensional nature of the manifold, characterizing its topology for interpolation is difficult. To overcome this problem, we apply a local linear embedding method to map the high dimensional manifold to a lower dimensional space, and perform the interpolation therein.[21]

Mapping the manifold to a locally linear lower dimensional space enables us to locally approximate the topology of the manifold with the Delaunay triangulation.[22] The local structure of the manifold by interpolating it on the d-simplex and then map it back to the high dimensional space: First, a d-simplex, along with its associated deformation field, is randomly selected from the manifold, and a d-dimensional random vector $r \in \mathbb{R}^{\mathrm{d}}$ is generated such that each of r's components are uniformly distributed on [0,1]. r is then normalized so that $\left\| r \right\|_1 = 1$. After that, a new deformation field can then be generated as $D^{*} = \sum_{i=1}^{d} r_i D_i$. Similar to the k parameter above, larger d results in more deformation fields contributing in the generation of the new one, which usually cause the new deformation to be smoother. The above process can be repeated infinitely, and allows us to create arbitrarily deformation fields.

It is noted that learning the object manifold and generating new objects are topics having been studied by many researchers[19,20,23,24] and we are not claiming the proposed algorithm being superior to any of the existing ones. In fact, our key objective is to simulate a known deformation so that we can always keep track of the exact boundary of the nucleus, for the ultimate purpose of evaluating nuclear segmentation algorithms.

The final nucleus is determined as $D^{*} \bigcirc P$. Moreover, the segmentation of the nucleus in it is also known, which is characterized by $D^{*} \bigcirc Q$. This is the key that enables us in using such synthesized images for evaluating nucleus segmentation.

2.1.2 Cytoplasm and Stroma—

Once we have generated the nuclear regions, we can “fill in” the cytoplasmic and the inter-celluar regions. While the segmentation of the cytoplasm is not available in our training data and in many times an ambiguous problem even for human eyes, it fortunately does not affect the present work significantly because we are mainly interested in the nuclei. In this work, we model the image content in the non-nuclear regions as a Markov Random Field (MRF).[25,26]

More explicitly, we start with any pixel location x in the newly synthesized image U that is not in, but bordering, the nucleus. The choice of such a location is because with some nucleus structures around, it is easier to infer the image intensity at this location. Then, a neighborhood ω_x is extracted centering at x. According to the assumption, some pixels in ω_x have already been known. Then, we find the most similar patch α in non-nuclear region of the training images. In the computation of similarity, the mean-square-difference (MSD) is computed over color image values of the already filled pixels. In addition, denote the similar patches as the set Γ:

$$\Gamma: = \left\{ \beta : MSD(\beta, U(\omega_x)) \le 1.3 MSD(\alpha, U(\omega_x)) \right\} \quad (2)$$

Based on the MRF assumption, the distribution of the center pixels of all the patches in is the same as that of the $U(x)$, weighted (inversely) by $MSD(\beta, U(\omega_x))$. A random sample is therefore drawn from this distribution and assigned to $U(x)$.

After this, we move onto the next un-filled pixel and repeat the same procedure above, until all the pixels in the image have been filled.

2.2 Evaluation of Segmentation Algorithm

The proposed algorithm is able to synthesize arbitrarily large image with known segmentation of nuclei in it. With that, we can evaluate the nucleus extraction algorithm at large scale.

He we provide a brief description of our segmentation algorithm to be evaluated. Given a new H&E stained digital pathology image $I: r \in \mathbb{R}^2 \to \mathrm{C}$, we want to extract the nuclei from it. To that end, we first normalize the color distribution in the CIE lab color space to correct for possible staining, imaging, and illumination artifacts. Then, the RGB image is separated into hematoxylin and eosin channels.[27]

Then, in the hematoxylin channel, the Otsu threshold is computed to give the initial extraction of the nuclei. This is followed by the fine tuning of the local statistics driven level set evolution.[28] The resulting segmentation may consist regions where several nuclei are clumped together. In order to de-clump the region and obtain the definition of each individual nuclei, the mean shift algorithm is used,[29,30] which gives the final segmentation of individual nuclei.

The algorithm is implemented using the Insight Toolkit[31] and the OpenSlide library,[32] which is able to run on large tiff-like images output from the microscopy scanner.

3. RESULTS

3.1 Image Synthesis Results

Figure 1 demonstrates the synthesized brain pathology image results. It can be observed that in addition to being visually realistic, the most desired feature is that the manual segmentation results have been fully transferred to the newly generated images, including the “touching” nuclei are also labeled correctly. Separating those nuclei are in particular important and challenging for the nucleus segmentation algorithm.

3.2 Quantitative Evaluation Results

Performing nucleus segmentation in the above images gives the segmentation results as shown in Figure 2. It can be observed that in the left-most image, the two touching nuclei are correctly separated. While in the next image, the touching nuclei in the top-right corner are not separated correctly, forming a Mickey-mouse head shape. The two nuclei below Mickey-mouse are not detected as two nuclei at all. Such difficulty is expected when we generate the images.

More importantly, the purpose of the proposed synthesis framework is to enable *large scale* evaluation. To that end, 1000 images containing a total of half-million ground truth nuclei are synthesized. The segmentation algorithm is tested on those images and an average Dice coefficient of 0.71 is achieved with standard deviation of 0.04.

4. CONCLUSIONS AND DISCUSSION

In this work, we proposed a method to synthesize arbitrarily large histopathology images from a small set of training images in which the nuclei have already been segmented. By the new and breakthrough work, we can systematically generate large validated data set to evaluate the automatic nucleus extraction algorithms.

There are several limitations that are being researched on. First, in the training image we do not have cytoplasm information marked out. Therefore, in the synthesized images, while we do observe the texture information in the non-nuclear region, we do not have clear definition for cytoplasm and cell boundary. This is acceptable for the current nucleus segmentation evaluation purpose. But will be extended for future more comprehensive study.

Second, the global/topological information is not well presented. We present in the result section the images generated from brain slides. However, for tissues from other organs, meso-scale features such as crypts and glands are also prominent features. Since our synthesis algorithm works at the fine resolution only, the current framework does not generate such meso-scale structure. To perform multi-scale tissue synthesis is our on-going research.

References

1. Gurcan MN, Boucheron LE, Can A, Madabhushi A, Rajpoot NM, Yener B. Histopathological image analysis: A review. IEEE reviews in biomedical engineering. 2009; 2:147–171. [PubMed: 20671804]
2. Naik S, Doyle S, Feldman M, Tomaszewski J, Madabhushi A. MIAAB workshop. Citeseer; 2007. Gland segmentation and computerized gleason grading of prostate histology by integrating low-, high-level and domain specific information; 1–8.
3. Karvelis PS, Fotiadis DI, Georgiou I, Syrrou M. Engineering in Medicine and Biology Society, 2006. EMBS'06. 28th Annual International Conference of the IEEE. IEEE; 2006. A watershed based segmentation method for multispectral chromosome images classification; 3009–3012.
4. Petushi S, Garcia FU, Haber MM, Katsinis C, Tozeren A. Large-scale computations on histology images reveal grade-differentiating parameters for breast cancer. BMC medical imaging. 2006; 6(1): 1. [PubMed: 16630362]
5. Weaver DL, Krag DN, Manna EA, Ashikaga T, Harlow SP, Bauer KD. Comparison of pathologist-detected and automated computer-assisted image analysis detected sentinel lymph node micrometastases in breast cancer. Modern pathology. 2003; 16(11):1159–1163. [PubMed: 14614056]
6. Yang L, Meer P, Foran DJ. Unsupervised segmentation based on robust estimation and color active contour models. IEEE Transactions on Information Technology in Biomedicine. 2005; 9(3):475–486. [PubMed: 16167702]
7. Wählby C, Sintorn I-M, Erlandsson F, Borgefors G, Bengtsson E. Combining intensity, edge and shape information for 2d and 3d segmentation of cell nuclei in tissue sections. Journal of Microscopy. 2004; 215(1):67–76. [PubMed: 15230877]

8. Naik S, Doyle S, Agner S, Madabhushi A, Feldman M, Tomaszewski J. 2008 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro. IEEE; 2008. Automated gland and nuclei segmentation for grading of prostate and breast cancer histopathology; 284–287.

9. Gurcan MN, Pan T, Shimada H, Saltz J. Engineering in Medicine and Biology Society, 2006. EMBS'06. 28th Annual International Conference of the IEEE. IEEE; 2006. Image analysis for neuroblastoma classification: segmentation of cell nuclei; 4844–4847.

10. Cooper LA, Kong J, Gutman DA, Wang F, Gao J, Appin C, Cholleti S, Pan T, Sharma A, Scarpace L, et al. Integrated morphologic analysis for the identification and characterization of disease subtypes. Journal of the American Medical Informatics Association. 2012; 19(2):317–323. [PubMed: 22278382]

11. Kong J, Cooper LA, Wang F, Gao J, Teodoro G, Scarpace L, Mikkelsen T, Schniederjan MJ, Moreno CS, Saltz JH, et al. Machine-based morphologic analysis of glioblastoma using whole-slide pathology images uncovers clinically relevant molecular correlates. PloS one. 2013; 8(11):e81049. [PubMed: 24236209]

12. Veta M, van Diest PJ, Kornegoor R, Huisman A, Viergever MA, Pluim JP. Automatic nuclei segmentation in h&e stained breast cancer histopathology images. PloS one. 2013; 8(7):e70221. [PubMed: 23922958]

13. Botev ZI, Grotowski JF, Kroese DP, et al. Kernel density estimation via diffusion. The Annals of Statistics. 2010; 38(5):2916–2957.

14. Devroye L. Proceedings of the 18th conference on Winter simulation. ACM; 1986. Sample-based non-uniform random variate generation; 260–265.

15. Gao Y, Zhu LJ, Bouix S, Tannenbaum A. SPIE Medical Imaging. International Society for Optics and Photonics; 2014. Interpolation of longitudinal shape and image data via optimal mass transport; 90342X–90342X.

16. Haber E, Rehman T, Tannenbaum A. An efficient numerical method for the solution of the l_2 optimal mass transfer problem. SIAM Journal on Scientific Computing. 2010; 32(1):197–211. [PubMed: 21278828]

17. Burkard RE, Dell'Amico M, Martello S. Assignment Problems, Revised Reprint. Siam; 2009.

18. Boyd S, Vandenberghe L. Convex optimization. Cambridge university press; 2004.

19. Joshi SC, Miller MI, Grenander U. On the geometry and shape of brain sub-manifolds. IJPRAI. 1997; 11(8):1317–1343.

20. Dambreville S, Rathi Y, Tannenbaum A. A framework for image segmentation using shape models and kernel space shape priors. IEEE transactions on pattern analysis and machine intelligence. 2008; 30(8):1385–1399. [PubMed: 18566493]

21. Roweis ST, Saul LK. Nonlinear dimensionality reduction by locally linear embedding. Science. 2000; 290(5500):2323–2326. [PubMed: 11125150]

22. Cormen TH, Leiserson CE, Rivest RL, Stein C. Introduction to algorithms. Vol. 6. MIT press; Cambridge: 2001.

23. Kendall DG. Shape manifolds, procrustean metrics, and complex projective spaces. Bulletin of the London Mathematical Society. 1984; 16(2):81–121.

24. Siddiqi K, Pizer S. Medial representations: mathematics, algorithms and applications. Vol. 37. Springer Science & Business Media; 2008.

25. Blake A, Kohli P, Rother C. Markov random fields for vision and image processing. Mit Press; 2011.

26. Efros AA, Leung TK. Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on. Vol. 2. IEEE; 1999. Texture synthesis by non-parametric sampling; 1033–1038.

27. Ruifrok AC, Johnston DA, et al. Quantification of histochemical staining by color deconvolution. Analytical and quantitative cytology and histology. 2001; 23(4):291–299. [PubMed: 11531144]

28. Lankton S, Tannenbaum A. Localizing region-based active contours. IEEE transactions on image processing. 2008; 17(11):2029–2039. [PubMed: 18854247]

29. Cheng Y. Mean shift, mode seeking, and clustering. IEEE transactions on pattern analysis and machine intelligence. 1995; 17(8):790–799.

30. Comaniciu D, Meer P. Mean shift: A robust approach toward feature space analysis. IEEE Transactions on pattern analysis and machine intelligence. 2002; 24(5):603–619.

31. Ibanez L, Schroeder W, Ng L, Cates J. The itk software guide. 2005

32. Goode A, Gilbert B, Harkes J, Jukic D, Satyanarayanan M, et al. Openslide: A vendor-neutral software foundation for digital pathology. Journal of pathology informatics. 2013; 4(1):27. [PubMed: 24244884]

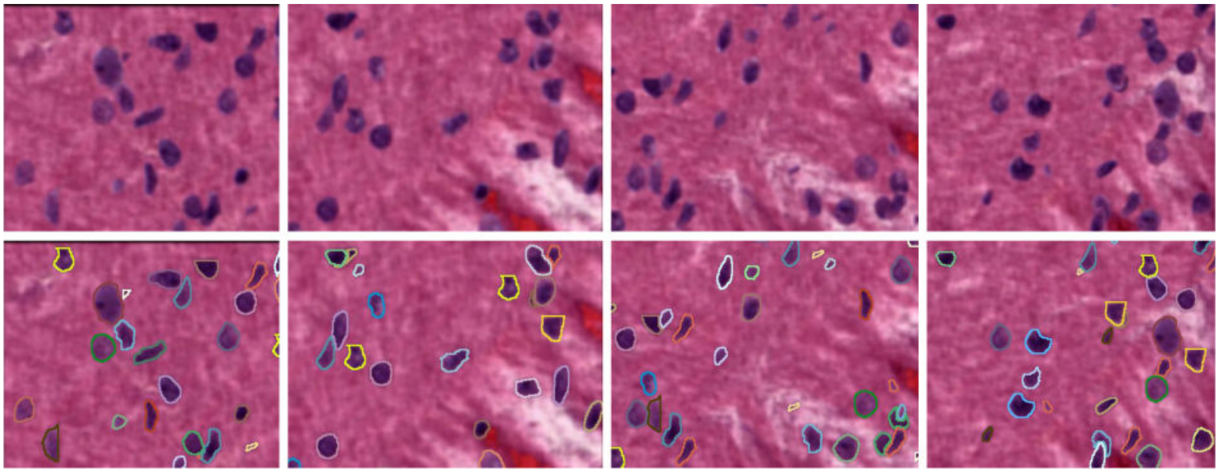

Figure 1.

Synthesized images with the ground truth segmentation. Note the “touching” nuclei: the proposed synthesis algorithm naturally handles the labeling of the touching nuclei. This enables the evaluation of the segmentation algorithms ability in separating them, which is often one of the most challenging step in the nucleus segmentation. However, the heavy occlusion between two pairs of nuclei in, e.g. the top-right corner of the second (from left) figure, should be very difficult, if possible at all, for the algorithm to separate.

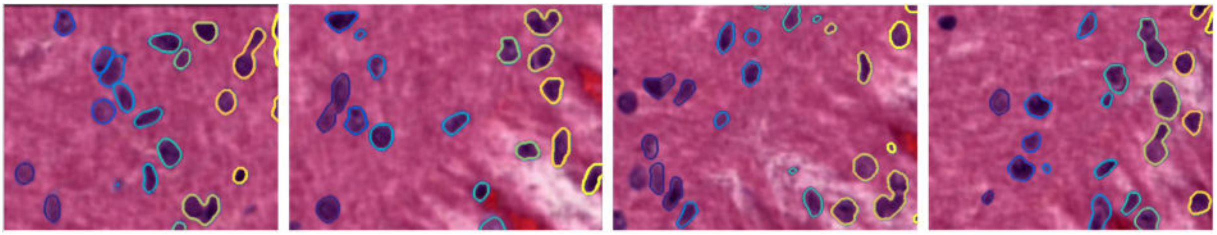

Figure 2.

The segmentation algorithm run on the above synthesized images. In the second (from left) figure, the touching nuclei in the top-right corner are not separated correctly, forming a Mickey-mouse head shape. The two nuclei below Mickey-mouse are not detected as two nuclei at all.