A Novel Segmentation Framework Using Sparse Random Feature in Histology Images of Colon Cancer

Kun Zhang[1], Huiyu Zhou[2], Li Chen[3], Minrui Fei[4], Jianguo Wu[1], and Peijian Zhang[1(✉)]

[1] School of Electrical Engineering, Nantong University, Nantong, China
zhangkun_nt@163.com, {wu.jg,zhang.pj}@ntu.edu.cn
[2] School of Electronics, Electrical Engineering and Computer Science, Queen's University, Belfast, UK
h.zhou@ecit.qub.ac.uk
[3] Department of Pathology, Medical College, Nantong University, Nantong, China
bll@ntu.edu.cn
[4] School of Mechatronic Engineering and Automaton, Shanghai University, Shanghai, China
mrfei@staff.shu.edu.cn

Abstract. In this paper, we present a novel segmentation framework for glandular structures in Hematoxylin and Eosin stained histology images, choosing poorly differentiated colon tissue as an example. The proposed framework' target is to identify precise epithelial nuclei objects. We start with staining separate to detect all nuclei objects, and deploy multi-resolution morphology operation to map the initial epithelial nuclei positions. We proposed a new bag of words scheme using sparse random feature to classify epithelial nuclei and stroma nuclei objects to adjust the rest nuclei positions. Finally, we can use the boundary of optimized epithelial nuclei objects to segment the glandular structure.

Keywords: Poorly differentiated glandular · Colon tissue · Color-deconvolution · Multi-resolution morphology · Sparse random feature

1 Introduction

Motivation: In this paper, we address the problem of segmenting challenging glandular structures in colon histology images. Glandular structures are important for the diagnosis of several epithelial cancers. Glands in epithelial tissue, which are difficult to be differentiated from other tissues, normally consist of lumen structure surrounded by epithelial nuclei objects at the boundary, which can be used as a strong cue for the extraction of glandular structures [1, 2]. This problem becomes more significant while there are other tissue constituents, such as stroma nuclei and cytoplasm, around the glands.

Related Work: Existing methods for glandular structure segmentation can be categorised into texture and structure based method. In texture approaches, Farjam et al. [3] used Gaussian filters to extract texture features from glandular structures. For structural approaches, Naik et al. [4] used a level set method to segment lumen areas in a gland. Nguyen et al. [5] employed the prior knowledge about glandular constituents in order to extract glandular regions. Gunduz-Demir et al. [6] proposed object graphs for the segmentation of glandular structures.

Contributions: In this work, we propose a novel framework to segment glandular structures in Hematoxylin and Eosin (H&E) stained histology images, choosing colon tissue as an example. The proposed framework starts with nuclei identification. We deploy a color-deconvolution method to find the nuclei position. Then, a multi-resolution morphology operation is investigated to interpret the epithelial nuclei spatial distribution. In this study, we assume that epithelial nucleus cannot be fully separated from stromal nucleus. Therefore, we deploy a sparse random features based Bag of words model to classify these nucleus. Finally, we use the boundary of epithelial nuclei for the final segmentation.

2 Sparse Random Matrix Optimization

Recently, random projection feature has shown promise for complex classification [7, 8]. It uses random matrix to project high level features to low dimension with the promise of core features preserved. A simple example is illustrated in Fig. 2, which shows the reconstruction of an ideal texture map based on random projection. The reconstruction results from different dimensions of projection are shown in Fig. 2(b), (c) and (d), respectively. With random projection, the original ideal texture is well reconstructed.

Based on the principle of distance preservation [9], Gaussian random matrix [9] and sparse random matrices [10, 11] have been sequentially proposed for random projection. Although significant progress has been made, the Gaussian random matrix and sparse random matrices still have prominent limitations: Gaussian random matrix causes highly computing for its dense distribution and it is difficult to construct Gaussian distribution via hardware, on the other hand the sparser matrix tends to yield weaker distance preservation. It has been proven that the irrelevance of matrices column vector is highly related to the projecting performance. In this paper, we use the angle between matrix neighbor column vectors to indicate the projecting performance. The first row of Table 1 is the minimum and maximum angles of the matrix, and the second row is the difference value of the minimum and maximum angles. The difference value is smaller the projecting performance is better. From Table 1, the random matrix and sparse random matrix are better than Gaussian random matrix.

Table 1. Parameters comparison under three random matrices

Gaussian random matrix	Sparse random matrix	Very sparse random matrix
71.42/107.34	81.04/100.93	82.04/110.93
35.92	19.89	28.89

Furthermore, we propose a new method to optimize sparse random matrix. Firstly, we build an empty matrix U, the length of U is λk, is k signal length and λ is positive integer. Secondly, we generate uniform random value to U from the range of 0 and 1. Thirdly, we recursively scan every value, if the value is larger than $5/6$, the value is reset as "1.732", if the value is smaller than $1/6$, the value is reset as "–1.732", other values are reset as "0". Fourthly, calculate the angle column by column, if the angle is small than threshold, assemble the corresponding column to a new $d \times k$ matrix. Specifically, the optimize algorithm is to form a new matrix by selecting the column vectors which angle is small than threshold. The new matrix not only satisfy RIP theorem but also to achieve irrelevance of matrices column vector.

Figure 3 shows the signal recover ability based on six random matrices. All optimization method based curves are the right side of the original method based curves, which means the signal recover ability of optimization methods are more useful than basic ones. Meanwhile, the ability of Gaussian and sparse random matrices are almost the same. To achieve the convenience of hardware design, we use optimization sparse random matrix in this paper.

3 Poorly Differentiated Glandular Segmentation Framework

Figure 1 summarizes the poorly differentiated glandular segmentation framework. Given an H&E image (Fig. 1a), nuclei locations are represented by color-deconvolution. (Figure 1b), necessary to perform morphology operation and random feature based classification to generate glandular contour (Fig. 1c).

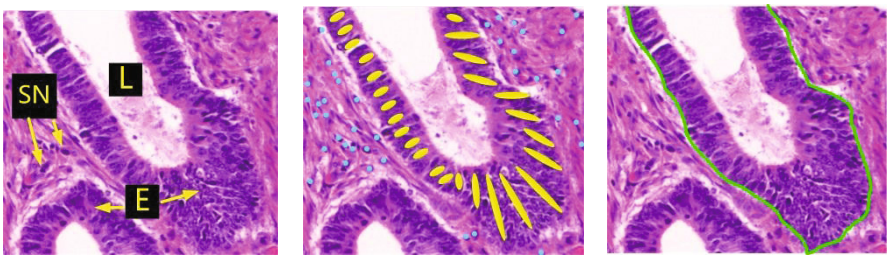
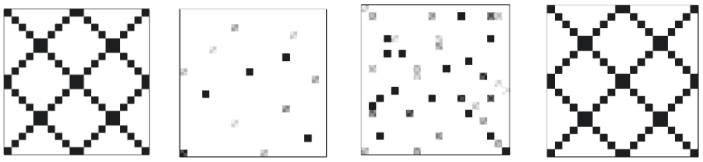


Fig. 1. A sample colon histology image showing various components (epithelial nucleus or E, stromal nucleus or SN, lumen or L)

3.1 Nuclei Identification

We employ a color-deconvolution method [12] to extract the Hematoxylin channel from the image. By thresholding the Hematoxylin channel using Otsu's threshold [13], we obtain a binary image corresponding to the approximate locations of nuclei in the image (Fig. 4).



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

Fig. 2. Random projection based ideal texture reconstruction: (a) original 20 × 20 ideal texture image, (b) reconstruction using 50 RP impact factors, (c) reconstruction using 100 RP impact factors, (d) reconstruction using 150 RP impact factors.

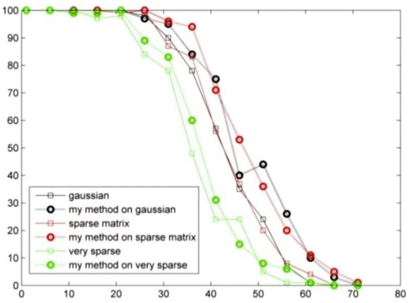
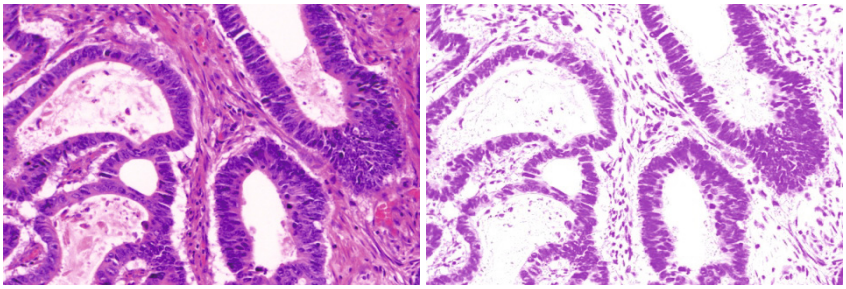


Fig. 3. The signal recover ability based on six random matrices.



(a) (b)

Fig. 4. (a) input image, (b) staining separated image using color-deconvolution.

3.2 Epithelial Layer Determination

Epithelial layers are most often formed by thick and solid objects such as epithelial nucleus. Therefore, a map describing the solidity and the connectedness of nuclei objects is calculated using this formula :

$$S_{map} = \frac{1}{R_s \times s_{\max}} \sum_{r=1}^{R_s} \sum_{s=1}^{s_{\max}} J_s((O_r(I_N))) \tag{1}$$

Where $O_r(X)$ is the morphological opening of image X with a disk of radius r, J_s is a removal operator of connected components with size smaller than s. R_s is the maximal radius and $s_{\max}$ the maximal connected component size. The use of these operations at different levels, i.e. with different values of r and s, agrees the variation in size of epithelial layers. The obtained map shows epithelial layers objects with higher solidity values than connective tissue nuclei objects (Fig. 5).

3.3 Epithelial Layer Optimization

Within the framework of RP based coding histogram coefficients in Fig. 6, for each RIRS feature vector [14] of the image, we can use RP to produce a compact and sparse representation, where only the entries corresponding to the same textons will have non-zero values, while the other entries in a compact vector are zeros. Then we can form the histogram coefficient as a new feature.

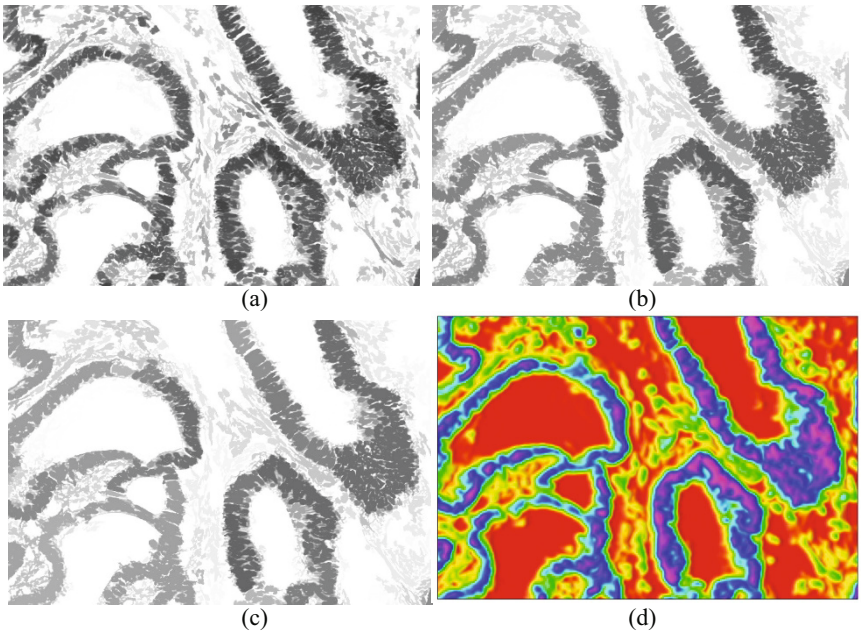


Fig. 5. (a)–(c) Epithelial layers identification after multi-resolution morphology operation, (d) Nuclei objects retained are overlayed in blue (Color figure online).

Figure 7 shows two nucleus of different classes (epithelial nuclei and stroma nuclei) and their 400 dimension histogram features. We can see that the histogram features of different classes are very different. For instance, in the upper two histograms, the features around “50”, “100” and “350” show strong differences and it is convenient for the later classification. At the same time, when texton number changes, the two histograms for the same image patch also show a sufficient difference. The upper

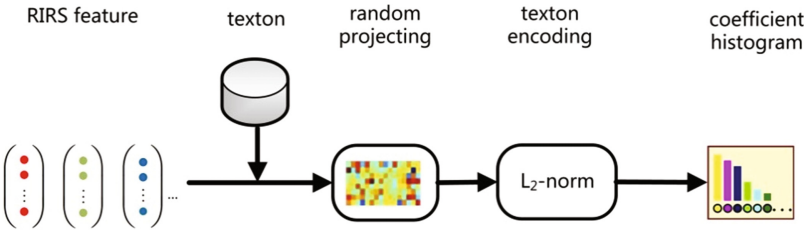


Fig. 6. Random projection based coding histogram coefficients. Best viewed in colour.

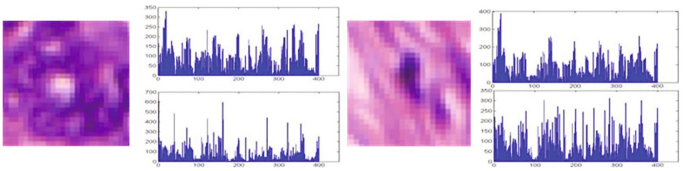


Fig. 7. epithelial nuclei and stroma nuclei images and their encoding histogram features (upper histogram – 9 textons and lower histogram – 25 textons). Best viewed in colour (Color figure online).

histograms are built from 9 textons and the lower histograms are constructed from 25 textons, and it is clear that the similarity of the two histograms based on 25 texton is much lower than 9 textons based histograms, and it is easy to find more discriminating features for classification.

4 Results and Discussion

To assess the effectiveness of our approach, the experiment was conducted on a public dataset of poorly differentiated H&E glandular tissue sections. The dataset, provided by Warwick university hospitals in the context of Gland Segmentation Challenge [15].

To quantitatively evaluate our segmentation results, the true positive (TP), false positive (FP), and false negative (FN) pixels are calculated. We use two quantitative criterions for the evaluation of our segmentation algorithm on the dataset. Table 2 shows the comparative quantitative performance of the proposed approach against 4 other methods in the literature; active contour [16], object-graphs [17], Morphology [18], and RIRS-bow [14], based on 3 quantitative measures: the precision also referred

Table 2. Quantitative evaluation of the segmentation methods.

Method	PPV	TRP	ACC
Active contour [16]	0.85	0.80	0.73
Object-graph [17]	0.91	0.87	0.77
Morphology [18]	0.89	0.86	0.78
RIRS-bow [14]	0.93	0.89	0.75
Proposed	0.85	0.87	0.81

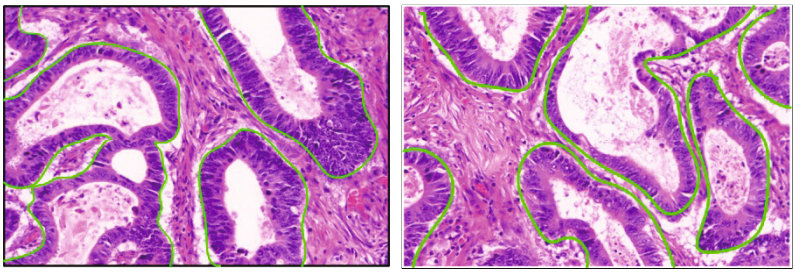


Fig. 8. Examples of segmentation results: image from Warwick university dataset. Best viewed in colour (Color figure online).

to as positive predictive value PPV = $\frac{TP}{TP + FP}$, the sensitivity referred to as the true positive rate TPR = $\frac{TP}{TP + FN}$, and the accuracy ACC = $\frac{TP}{TP + FP + FN}$. Figure 8 is examples of segmentation results.

5 Conclusion

This work presents a new gland segmentation approach based on gland structural feature. This method begin with staining separation to detect nuclei objects, then advance multi-resolution morphological operator is applied to map the initial epithelial nuclei objects. After that, a developed sparse random feature is used to classify the rest nuclei objects from epithelial nuclei class and stroma nuclei class. Finally, the boundary of epithelial nuclei objects is the core clue to segment the gland.

Acknowledgements. This work was financially supported by the Natural Science Foundation of Jiangsu Province, China under grant No. BK20170443. Nantong Research Program of Application Foundation under Grant No. GY12016022 and Dr. H Zhou is currently supported by UK EPSRC under Grant EP/N011074/1, and Newton Advanced Fellowship under Grant NA160342.

References

1. Gleason, D.F., Mellinger, G.T.: Prediction of prognosis for prostatic adenocarcinoma by combined histological grading and clinical staging. J. Urol. **167**(2), 953–958 (2002)
2. Hamilton, S.R., Aaltonen, L.A.: International Agency for Research on Cancer, World Health Organization, and others, Pathology and genetics of tumours of the digestive system (2000)
3. Farjam, R., Soltanian-Zadeh, H., Jafari-Khouzani, K., Zoroofi, R.A.: An image analysis approach for automatic malignancy determination of prostate pathological images. Cytometry Part B Clin. Cytometry **72**(4), 227–240 (2007)
4. Naik, S., Doyle, S., Madabhushi, A., Tomaszewski, J., Feldman, M.: Gland segmentation and gleason grading of prostate histology by integrating low-, high-level and domain specific information. In: Workshop on Microscopic Image Analysis with Applications in Biology (2007)

5. Nguyen, K., Jain, A.K., Allen, R.L.: Automated gland segmentation and classification for gleason grading of prostate tissue images. In: 2010 20th International Conference on Pattern Recognition (ICPR), pp. 1497–1500. IEEE (2010)
6. Gunduz-Demir, C., Kandemir, M., Tosun, A.B., Sokmensuer, C.: Automatic segmentation of colon glands using objectgraphs. Med. Image Anal. **14**(1), 1–12 (2010)
7. Varma, M., Zisserman, A.: A statistical approach to material classification using image patches. IEEE Trans. Pattern Anal. Mach. Intell. **31**(11), 2032–2047 (2009)
8. Zhang, J., Marszalek, M., Lazebnik, S., Schmid, C.: Local features and kernels for classification of texture and object categories: a comprehensive study. Int. J. Comput. Vis. **73** (2), 213–238 (2007)
9. Liu, L., Fieguth, P.: Texture classification from random features. IEEE Trans. Pattern Anal. Mach. Intell. **34**(3), 574–586 (2012)
10. Indyk, P.: Sparse recovery using sparse random matrices. In: López-Ortiz, A. (ed.) LATIN 2010. LNCS, vol. 6034, p. 157. Springer, Heidelberg (2010). doi:10.1007/978-3-642-12200-2_15
11. Lu, W., Li, W., Kpalma, K., et al.: Sparse matrix-based random projection for classification. Comput. Sci. **12**, 581–607 (2013)
12. Macenko, M., Niethammer, M., Marron, J., Borland, D., Woosley, J.T., Guan, X., Schmitt, C., Thomas, N.E.: A method for normalizing histology slides for quantitative analysis. In: International Symposium on Biomedical Imaging, vol. 9, pp. 1107–1110 (2009)
13. Otsu, N.: A threshold selection method from gray-level histograms. IEEE Trans. Syst. Man Cybern. **9**(1), 62–66 (1979)
14. Zhang, K., Crookes, D., Diamond, J., Fei, M., Wu, J., Zhang, P., Zhou, H.: Multi-scale colorectal tumour segmentation using a novel coarse to fine strategy. In: British Machine Vision Conference (BMVC) (2016)
15. Sirinukunwattana, K., Pluim, J.P., Chen, H., et al.: Gland segmentation in colon histology images: the glas challenge contest. Med. Image Anal. **35**, 489 (2016)
16. Cohen, A., Rivlin, E., Shimshoni, I., Sabo, E.: Memory based active contour algorithm using pixel-level classified images for colon crypt segmentation. Comput. Med. Imaging Graph. **43**, 150–164 (2015)
17. Gunduz-Demir, C., Kandemir, M., Tosun, A., Sokmensuer, C.: Automatic segmentation of colon glands using object-graphs. Med. Image Anal. **14**(1), 112 (2010)
18. Cheikh, B.B., Bertheau, P., Racoceanu, D.: A structure-based approach for colon gland segmentation in digital pathology. In: SPIE Medical Imaging, vol. 9791, p. 97910J (2016)