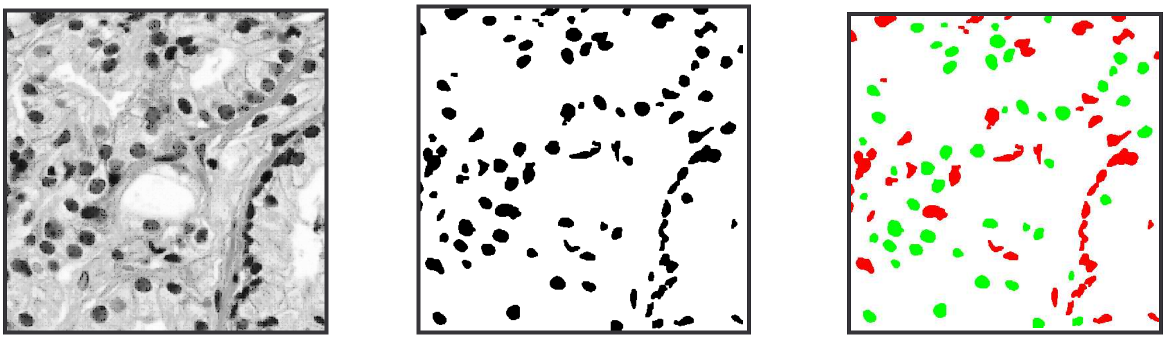
Detection of Nuclei by Unsupervised Manifold Learning

Muhammad Arif[a] and Nasir Rajpoot[b]

[a] Pakistan Institute of Engineering and Applied Sciences, Nilore, Islamabad, Pakistan
[b] Department of Computer Science, University of Warwick, Coventry, UK

Abstract: Shape related signatures of nuclei in a tissue section are important for diagnosis and prognosis of cancer. Understandably, the process of demarcation of nuclei for cytometry with high degree of confidence is the most difficult part as the tissue section is fraught with staining artifacts and frequently contains other objects such as overlapping nuclei, nuclear debris, and extracellular structures. In this paper, we address this problem using a novel clustering algorithm for various shapes in prostate histopathology images using an unsupervised manifold learning paradigm. Experimental results with two-dimensional embedding of the shapes using diffusion maps demonstrate that various shapes in the tissue section are organized in accordance to the degree of complexity of their boundaries. This important observation can be exploited in the development of computerized techniques for image based cytometry.



(a) (b) (c)

Figure 1: (a) A typical prostate tissue specimen (b) Binary image showing potential nuclei, overlapping nuclei, nuclear debris, and extracelluar structures (c) Nuclei detected (shown in green) by our detection algorithm; the objects shown in red are the ones rejected by the algorithm.

1 Introduction

Today cancer is one of the major causes of deaths throughout the world. Mortality rate can be reduced if it is diagnosed at an early stage. Traditionally, cancer and its level of malignancy (i.e., grading) are detected by pathologists who study microscopic images of tissues removed from the patients. For over a decade now, there has been a considerable amount of interest in research and development of computer-assisted systems for diagnosis and prognosis of cancer, due to the following reasons:

1. The judgment of a pathologist is subjective and significant inter- and intra-observer variability and poor reproducibility has been reported [1, 2]. Analyzing the tissues quantitatively using machine vision techniques can reduce this variability.
2. In the very early stages of the malignancy, the pathologist may be unable to detect the changes in the tissues. By introducing objectivity in the analysis, these otherwise undetectable changes can be detected [3].
3. A significant number of imaging modalities have been developed in recent years. To increase the reliability of judgment, it would be more judicious to fuse the information extracted from the huge amount of data generated by these imaging modalities as well as from the computer based analysis of tissue specimens.
4. The task of analyzing histopathological data by human experts is dull, tedious and laborious. Machines are well suited for such chores with high throughput. Hence nature of the job also demands automation of the process or at least development of ancillary systems to decrease the workload for pathologists.

One of the main areas of research into quantitative histopathology for computer-assisted diagnosis systems has been to quantify changes at one of the two levels: tissue level and cellular level [4]. Computerized methods at both levels extract diagnostically important features from stained tissues and employ statistical and machine learning methods to detect any abnormality. Most of the techniques at tissue level rely on the texturedness of the tissue and on the spatial arrangement of cells in histological sections [5, 6, 7, 8, 9, 10]. The second class of methods exploits the morphological features of cells and nuclei [11,14,15], such as quantitative descriptors of nuclei including area,

E-mail addresses: {arif,nasir}@dcs.warwick.ac.uk

radius, major/minor axis, compactness, perimeter, fractal dimension and texture of the chromatin inside the nucleus, to name a few [13].

Nuclei are particularly very important objects in the tissue to be analyzed for the development of computer-assisted systems as they can be easily differentiated from rest of the structural components by selective staining methods. The success of nuclear morphometry based techniques depends on how reliably the nuclei have been recognized in a histological section fraught with overlapping nuclei, nuclear debris, cellular and extra cellular structures, staining and image acquisition related artifacts.

Inspired by the seminal work of Coifman et al. on diffusion maps [16], we address the problem of separating various parts of the tissue in a stained specimen by posing it as an unsupervised manifold learning problem in a low-dimensional space and by using diffusion maps for manifold learning. Preliminary results are very encouraging and demonstrate that various objects in the tissue section are organized in the new representation (using only 2 dimensions) in accordance with the degree of complexity of their boundaries. The proposed detection algorithm is fast and only requires a thresholding operation, with the threshold being the same for all images that we experimented with, after low-dimensional embedding to separate the nuclei from a given histopathological image.

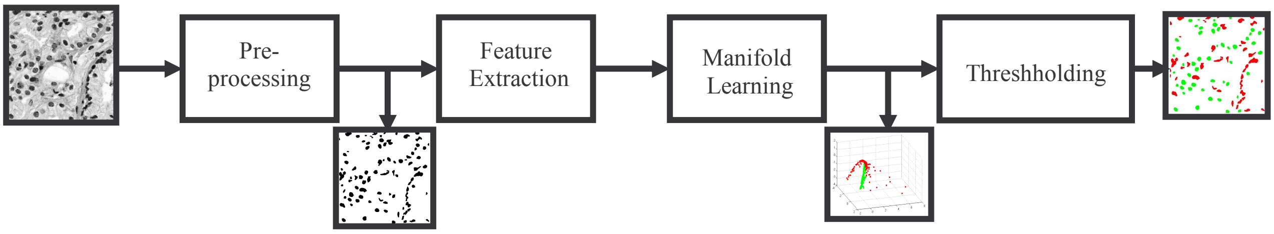


Figure 2. Block diagram of the proposed algorithm

2 Materials and Methods

We used Haematoxylin and Eosin (H&E) stained color images of prostate tissue samples for experimentation. All of the images were captured under similar conditions. The images were converted into gray scale and filtered using edged preserving smoothing filter to remove the noise. A typical specimen of the stained tissues is shown in the Figure 1 (a).

2.1.1 Preprocessing and Feature Extraction

The general block diagram of the proposed algorithm is shown in Figure 2. The gray images were processed by an edge-preserving smoothing filter, namely *anisotropic diffusion*, to remove the noise. In a typical H&E stained image of a tissue section, cell nuclei are darker than the surrounding cytoplasm and extracelular matter. Exploiting this fact, the class of potential nuclei are segmented from the specimen using *k*-means clustering on the intensity values. The resultant image with expected nuclei is then converted into a binary image. Morphological operations are performed on the dichotomized image to smooth it and to remove segmentation artifacts and objects with very small areas. The resultant binary image is shown in Figure 1 (b).

As can be seen in Figure 1(b), the processes of segmentation and binarization do not isolate the nuclei completely from other objects such as custers of overlapping nuclei, nuclear debris, cellular and extracellular structures and staining artifacts. In order to learn the shape manifolds of nuclei and other objects in an unsupervised manner, we extract the boundaries of all objects. The extracted boundary points of each object are then resampled into an equal number of points N using cubic spline interpolation. A centroidal distance function r_i, for $i=1,2,\dots,N,$ is computed as follows [12],

$$r_i = \sqrt{([x_i - x_c]^2 + [y_i - y_c]^2)} \tag{1}$$

where (x_i,y_i) denote the coordinates of the ith boundary point of the object. The distance vector $\boldsymbol{r}=\{r_1,r_2,\dots r_N\}$ is transformed into frequency domain using FFT. Our feature vector $\boldsymbol{f}$ for the detection algorithm is derived as follows,

$$\boldsymbol{f} = \left[\frac{|F_1|}{|F_0|}, \frac{|F_2|}{|F_0|}, \dots, \frac{|F_{N/2}|}{|F_0|}\right]^T \tag{2}$$

where $|F_i|$ denote the ith Fourier coefficient, with F_0 being the DC component. In the above equation, taking the magnitude of the coefficients yields rotation invariance and their division by $|F_0|$ results in scale invariance.

2.1.2 Unsupervised Manifold Learning

To improve the performance of a classifier or a clustering algorithm, it is desirable to extract relevant features concealed in high dimensional data. Generally, the relevant features lie on a low-dimensional sub-manifold embedding of the high dimensional space. Many algorithms have been proposed to learn a suitable low-dimensional embedding from the given data. The classical techniques such as PCA, ICA, MDS etc. assume that the low-dimensional manifold is linear. In recent years, locality preserving dimensionality reduction techniques such as Locally Linear Embeddings [17], Laplacian Eigen Maps [18], Diffusion Maps [16] and Hessian Eigen Maps [19] have been proposed when the features are lying in a non-linear low-dimensional manifold. These methods exploit the spectral properties (eigenvectors and eigenvalues) of some kind of similarity matrices.

In this work, we employ a diffusion map based framework for unsupervised manifold learning as it clearly defines a robust (with respect to noise) distance metric on the data set reflecting its connectivity [20]. As a consequence, the number of Eigenvectors in low-dimensional embedding can be determined by time parameter in Markov chain. In other methods, it is not clear how to choose the number of eigenvectors in the embedding [20].

For a given set of feature vectors $\Omega=\{\boldsymbol{f_1}, \boldsymbol{f_2}, \ldots, \boldsymbol{f_n}\}$ corresponding to boundaries of various objects in the tissue section, the first step of diffusion maps based framework is to consider feature vectors as nodes of a symmetric graph and compute pairwise similarity matrix between points $\boldsymbol{f}_i$ and $\boldsymbol{f}_j$ using Gaussian Kernel with width ε,

$$w(\boldsymbol{f}_i, \boldsymbol{f}_j) = \exp\left(-\frac{\left\| \boldsymbol{f}_i - \boldsymbol{f}_j \right\|^2}{2\varepsilon} \right) \tag{3}$$

The next step is to define Markov Random Walk on the graph by defining the Markov matrix $\boldsymbol{P}$, with its (i,j)th element p_{ij} given as follows

$$p_{ij} = \frac{w(\boldsymbol{f}_i, \boldsymbol{f}_j)}{d(\boldsymbol{f}_i)} \tag{4}$$

where $d(\boldsymbol{f}_i)$ denotes the degree of node $\boldsymbol{f}_i$ in the graph and is defined as

$$d(\boldsymbol{f}_i) = \sum_{z \in \Omega} w(\boldsymbol{f}_i, \boldsymbol{z}). \tag{5}$$

The (i,j)th element of $\boldsymbol{P}^t$ (i.e., the tth iterate of $\boldsymbol{P}$) represent probability of going from node $\boldsymbol{f}_i$ to node $\boldsymbol{f}_j$ in t steps. If $\{\lambda_l\}$ is the sequence of eigenvalues of $\boldsymbol{P}$ such that $|\lambda_0| \geq |\lambda_1| \geq \ldots,$ and $\{\psi_l\}$ are the corresponding eigen functions, then a mapping from the data set Ω to a low-dimensional Euclidean space $\mathfrak{R}^m$,where m is the dimensionality of the lower-dimensional space and can be a function of t, is given by (see [16, 19] for details),

$$\Psi^t : \boldsymbol{f} \mapsto (\lambda_1^t \psi_1(\boldsymbol{f}), \lambda_2^t \psi_2(\boldsymbol{f}), \ldots. \lambda_m^t \psi_m(\boldsymbol{f}))^T \tag{6}$$

Spectral fall-off and the time t of the random walk are the main factors contributing to dimensionality reduction. Consequently, for large value of t, we will be able to capture large-scale structures in the data with fewer diffusion coordinates [21]. A natural consequence of this embedding is that it captures the intrinsic dimensionality of the high dimensional data which could be depicted by the way the objects are organized in the new, lower dimensional space.

The two-dimensional diffusion map for $(N/2\text{-}1)$-dimensional feature vectors (with $N = 100$ boundary points) for the tissue sample shown in Figure 1 (a) is shown in Figure 3 (a). In both figures, nuclei are shown in green and rest of the entities in red. These figures show that the low-dimensional embeddings have successfully captured in high dimensional data characterizing nuclei-like objects.

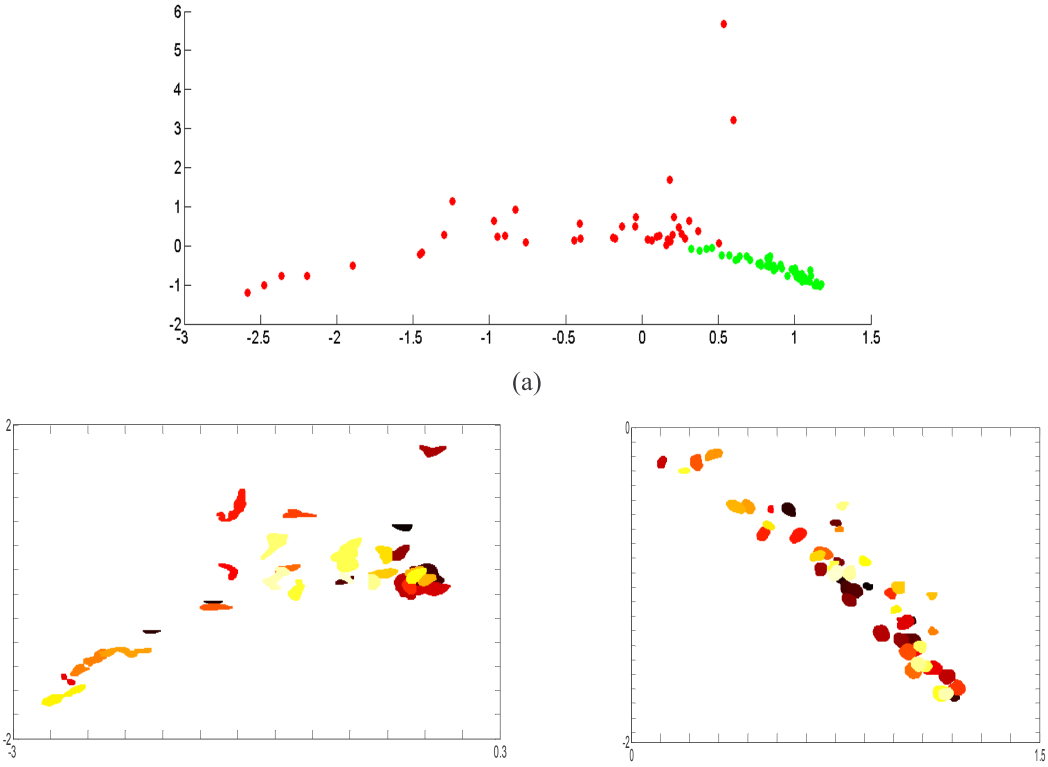
3 Results and Discussion

Diffusion coordinates were computed for high dimensional feature vector data corresponding to objects in the segmented tissue of Figure 1 (a) as outlined in the previous section. Figure 3 (a) shows the low-dimensional embedding of high dimensional original feature vectors extracted from the Figure 1 (b) in new space. To investigate for some interesting patterns emerging from the embedding, two non-overlapping sections of the Figure 3 (a) have been zoomed in (b) and (c) and the objects in the binarized image corresponding to Figure 1 (a) have been overlaid approximately in their respective coordinates in the embedding.

Visual examination of these plots demonstrates that as we move roughly from left to right and from bottom to top, the degree of complexity in the shape of objects increases. To put it another way, we may say that regular shaped nuclei have been clustered on the bottom right of the plot while non-regular shapes such as bunches of overlapping nuclei, possibly irregular nuclei, debris of nuclei, nuclei in different stages of mitosis, staining artifacts etc. have distinct regions in sub-manifold. The same pattern was observed when the experiment was repeated for several specimens.

Interestingly almost all the regular shaped nuclei can easily be isolated from rest of the objects by thresholding the values of the first two eigen vectors at zero crossings. The threshold values of zero are chosen due to the nature of the problem (binary clustering) using the spectral clustering theory [22]. Figure 1 (c) shows the tissue section with potential nuclei with green color by this technique. Figure 4 (c) also shows the results of successful classification of objects in the binary image by thresholding for another specimen shown in Figure 4 (a). Note that in Figure 3 (a) the coordinates of potential nuclei in diffusion maps recovered by mere thresholding have been coded in green and other objects in red.

Intuitively, these observations seem to be very promising in the field of quantitative cytometry. However, it has yet to be established, by extensive experimentation on ground-truth, how significant these findings are for diagnosis and prognosis of cancer.



(a)

(b)

(c)

Figure 3. (a) Diffusion maps for the binary image of Figure 1(b). (b) & (c) The objects in the binary image corresponding to the their embedded coordinates overlaid approximately on their respective coordinates in (a).

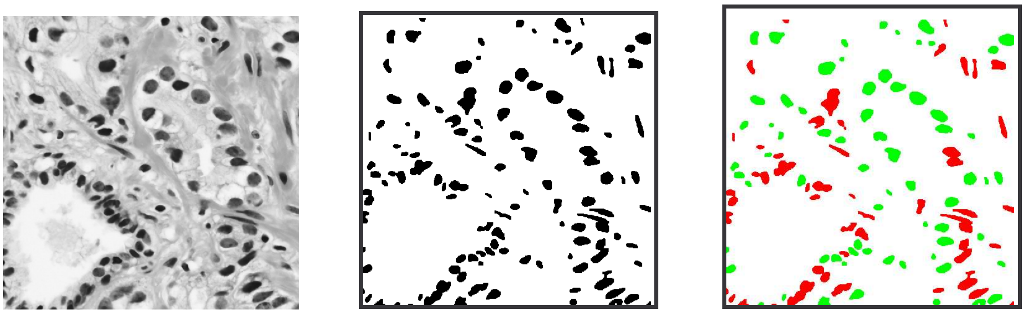


Figure 4: Another prostate tissue specimen (b) Binary image containing potential nuclei, overlapping nuclei, nuclear debris, and extracelluar structures (c) Potential Nuclei (shown in green) detected via shape manifold learning with diffusion maps.

Conclusions

The problem of detecting nuclei in a tissue section for image-based cytometry has been addressed in this paper using an unsupervised manifold learning paradigm. Preliminary results demonstrate that this technique successfully reorganizes various objects of the tissue section in accordance with the complexity of the objects' boundaries. The pattern that emerges in the diffusion space can be incorporated in the development of efficient and reliable systems for quantitative histopathology.

Acknowledgements

The first author would like to gratefully acknowledge the Higher Education Commission, Government of Pakistan, for financial support and the Department of Computer Science, University of Warwick, UK for hosting his visit and for providing facilities to carry out this research.

References

1. S.M. Ismail, A.B. Colclough, J.S. Dinnen et al., "Observer Variations in Histopathological Diagnosis and Grading of Cervical Intraepithelial Neoplasia," *Br. Med. J.*, 298(6675): 707-710, April 1989.
2. P.J Klinkhamer, G.P. Vooijs, & A.F. de Haan, "Intraobserver and Inerobserver Variability in the Diagnosis of Epithelial Abnormalities in Cervical Smears, *" Acta Cytology*, 32(6):794-800, Nov-Dec 1998.
3. P. Bartlets, "Computer-Generated Diagnosis and Image Analysis: An Overview" *Cancer Supplement*, Vol. 69, No. 6, March 15, 1992.
4. C. Demir & B. Yener, "Automated Cancer Diagnosis Based on Histopathological Images: A Systematic Survey," *Rensselaer Polytechnic Institute Technical Repor,t* TR-05-09, 2005.
5. C. Demir, S. Humayun Gultekin, & B. Yener, " Learning the Topological Properties of Brain Tumors," *IEEE/ACM Transactions on Computational Biology and Bioinformatics,* Vol. 2, No. 3, July-September 2005.
6. Qing Ji, John Engel, and eric Craine, " Texture Analysis for Classification of Cervix Lesions," *IEEE Transactions on Medical Imaging,* Vol. 19, No. 1, November 2000.
7. K. Masood, N. Rajpoot, K. Rajpoot et al., "Hyperspectral Texture Analysis for Colon Tissue Biopsy Classification," *International Symposium on Health Informatics and Bioinformatics*, 2007.
8. W. Christen-Barry & A. Partin, "Quantitative Grading of Tissue and Nuclei in Prostate Cancer for Prognosis Prediction," *John-Hopkins APL Technical Digest*, Vol. 18, Number 2, 1997.
9. H. Qureshi, N. Rajpoot, K. Masood et al., "Classification of Meningiomas using Discriminant Wavelet Packets and Learning Vector Quantization," *In Proceedings of Medical Image Understanding and Analysis*, 2006.
10. B. B. Chaudhuri, K. Rodenacker, & G. Burger, " Charactrization and Featuring of Histological Section Images," Pattern Recognition Letters, 7, pp. 245-262, 1988.
11. J. P. Thiran & B. Macq, " Morphological Feature Extraction for the Classification of Digital Images of Cancerous Tissue," *IEEE Transactions on Biomedical Engineering*, Vol. 43, No. 10, October 1996.
12. H. Kauppinen, T. seppanen, & M. Pietikainen, "An experimental Comparison of Autoregressive and Fourier-based Descriptors in 2-D Shape Classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. 17, pp. 201-207, 1995.
13. K. Rodenacker, & E. Bengtsson, "A Feature Set for Cytometry on Digitized Microscopic Images," *Analytical Cellular Pathology*, pp. 1-36, 25(2003), ISO Press.
14. W. H. Wolberg, W. N. Street, D. M. Heisey, & O. L. Mangasarian, "Computer-derived Nuclear Features Distinguish Malignant from Benign Breast Cytology," *Hum. Pathology*, 26(7):792-796, July 1995.
15. K. Yogesan, T. Jorgensen, F. albregtsen et al., "Entropy-Based Texture Analysis of chromatin Structure in Advanced Prostate Cancer," *Cytometry,* 24:268-276, 1996.
16. R. R. Coifman & S. Lafon, "Diffusion Maps," *Applied and Computational Harmonic Analysis*, Special Issue on diffusion maps and wavelets, Vol. 21, pp. 5-30, July 2006.
17. S. Roweis & L. Saul, "Nonlinear Dimensionality Reduction by Locally Linear Embedding," *Science*, vol. 290, pp. 2323-2326, 2000.
18. M. Bilkin & P.Niyogy, "Laplacian Eigen Maps for Dimensionality Reduction and Data Representation," *Neural Computation*, Vol. 6. No. 15, pp. 1373-1396, June 2003.
19. D. Donoho & C. Grimes, " Hessian Eigenmaps: New Locally Linear Embedding Techniques for High Dimensional Data," *Proc. Nat'l Academy of Sciences,* vol. 100, no. 10, pp. 5591-5596, May 2003.
20. S. Lafon & A. B. Lee, "Diffusion Maps and Coarse-Graining: A Unified Framework for Dimensionality Reduction, Graph Partitioning, and Data Set Parmeterization," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. 28, No. 9, September 2006.
21. S. Lafon, Y. Keller & R. R. Coifman, "Data Fusion and Multicue Data Matching by Diffusion Maps," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. 28, No. 11, November 2006.
22. J. Shi & J. Malik, "Normalized cuts and image segmentation," *IEEE. Trans. on Pattern Analysis and Machine Intelligence*, 22:888-905, 2000.