# Spatially Aware Cell Cluster(SpACCl) Graphs: Predicting Outcome in Oropharyngeal p16+ Tumors

Sahirzeeshan Ali[1], James Lewis[2], and Anant Madabhushi[1,⋆]

[1] Case Western University, Cleveland, OH USA
[2] Surgical Pathology, Washington University, St Louis, MO USA

**Abstract.** Quantitative measurements of spatial arrangement of nuclei in histopathology images for different cancers has been shown to have prognostic value. Traditionally, graph algorithms (with cell/nuclei as node) have been used to characterize the spatial arrangement of these cells. However, these graphs inherently extract only global features of cell or nuclear architecture and, therefore, important information at the local level may be left unexploited. Additionally, since the graph construction does not draw a distinction between nuclei in the stroma or epithelium, the graph edges often traverse the stromal and epithelial regions. In this paper, we present a new spatially aware cell cluster (SpACCl) graph that can efficiently and accurately model local nuclear interactions, separately within the stromal and epithelial regions alone. SpACCl is built locally on nodes that are defined on groups/clusters of nuclei rather than individual nuclei. Local nodes are connected with edges which have a certain probability of connectedness. The SpACCl graph allows for exploration of (a) contribution of nuclear arrangement within the stromal and epithelial regions separately and (b) combined contribution of stromal and epithelial nuclear architecture in predicting disease aggressiveness and patient outcome. In a cohort of 160 p16+ oropharyngeal tumors (141 non-progressors and 19 progressors), a support vector machine (SVM) classifier in conjunction with 7 graph features extracted from the SpACCl graph yielded a mean accuracy of over 90% with PPV of 89.4% in distinguishing between progressors and non-progressors. Our results suggest that (a) stromal nuclear architecture has a role to play in predicting disease aggressiveness and that (b) combining nuclear architectural contributions from the stromal and epithelial regions yields superior prognostic accuracy compared to individual contributions from stroma and epithelium alone.

⋆ Research reported in this publication was supported by the National Cancer Institute of the National Institutes of Health under Award Numbers R01CA136535-01, R01CA140772-01, R43EB015199-01, and R03CA143991-01. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health

K. Mori et al. (Eds.): MICCAI 2013, Part I, LNCS 8149, pp. 412–419, 2013.

## 1 Introduction

Graph theory has emerged as a popular method to characterize the structure of large complex networks leading to a better understanding of dynamic interactions that exist between their components [1]. Nodes with similar characteristics tend to cluster together and the pattern of this clustering provides information as to the shared properties, and therefore the function, of those individual nodes [4]. Despite their complex nature, cancerous cells tend to self-organize in clusters and exhibit architectural organization, an attribute which forms the basis of many cancers [2].

In the context of image analysis and digital pathology, some researchers have shown that spatial graphs and tessellations such as those obtained via the Voronoi (VT), Delaunay (DT), and minimum spanning tree (MST), built using nuclei as vertices may actually have biological context and thus be potentially predictive of disease severity [1, 3] . These graphs have been mined for quantitative features that have shown to be useful in the context of prostate and breast cancer grading [1]. However, these topological approaches focus only on local-edge connectivity. Moreover, these graphs inherently extract only global features and, therefore, important information involving local spatial interaction may be left unexploited. Additionally, since no distinction is made between the nuclear vertices lying in either the stroma or epithelium, the graph edges often traverse the stromal and epithelial regions (see Figure 1 and 2).

Until recently morphology of the stroma has been largely ignored in characterizing disease aggressiveness [7]. However, there has been recent interest in looking at possible interactions between stromal and epithelial regions and the role of this interaction in disease aggressiveness [6]. However, constructing global Voronoi and Delaunnay graphs which connect all the nuclei (involved in the stroma and epithelium) may not allow for capturing of local tumor heterogeneity. Additionally these global graph constructs do not allow for evaluating contributions of the stromal or epithelial regions alone. Consequently there is a need for locally connected graphs which are spatially aware which will allow for quantitative characterization of spatial interactions within the stroma and epithelial regions separately and hence for combining attributes from both the stromal and epithelial graphs.

Human papillomavirus-related (p16 positive) oropharyngeal squamous cell carcinoma (oSCC) represents a steadily increasing proportion of head and neck cancers and has a favorable prognosis [4]. However, approximately 10% of patients develop recurrent disease, mostly distant metastasis, and the remaining patients often have major morbidity from treatment. Hence, identifying patients with more aggressive (rather than indolent) tumor is critical. In this work we seek to develop an accuracy, image based predictor to identify new features in oSCC cancer, thereby providing new insights into the biological factors driving the progression of oSCC disease. This paper presents a new Spatially Aware Cell Cluster(SpACCl) graph that can efficiently and accurately model local nuclear architecture within the stromal and epithelial regions alone. The novel contributions of this work include,

1. Unlike global graphs (where the vertices are not spatially aware), SpACCl is built locally on nodes that are defined on groups/clusters of nuclei rather than individual nuclei. Consequently, SpACCl can be mined for local topological information, such as clustering and compactness of nodes (nuclei), which we argue may provide image biomarkers for distinguishing between indolent and progressive disease.
2. SpACCl allows for implicit construction of two separate graphs within each of the stromal and epithelial regions to extract features from both the regions. To distinguish epithelium nodes from stromal nodes to individually extract graph features from the two regions, a super-pixel based support vector machine classifier is employed to separate out the regions in the image into stromal and epithelial compartments. Stromal and epithelial interactions are explored by combining graph features extracted from the two regions and using these features to train a classifier to identify progressors (poor prognostic tumors) in p16+ oropharyngeal cancers.

## 2 Constructing Spatially-Aware Cell Cluster Graphs

The intuition behind SpACCl is to capture clustering patterns of nuclei in histologic tissue images and extracting topological properties and attributes that can quantify tumor morphology efficiently. Formally, SpACCl is defined as $G_i = (V_i, E_i)$, where $i \in \{\text{epithelium}, \text{stroma}\}$, $V_i$ and $E_i$ are the set of nodes and the edges respectively. Construction of SpACCl is illustrated in Figure 1 and described in detail below.

### 2.1 Distinguishing Stromal and Epithelial Compartments

The entire image is partitioned into small, spatially coherent cells known as super-pixels [8]. Nuclei within these super-pixels were identified by performing dendogram clustering of the mean intensity values (RGB) of each superpixel, within which we measured the intensity and texture (local binary patterns and harlick) of the superpixel and its neighbors. We then classified superpixels as being either within the epithelium or stroma by training a Support Vector Machine (SVM) classifier on these measurements with hand-labelled superpixels from 100 images.

### 2.2 Cluster Node Identification

The second step is to identify closely spaced (clusters) of nuclei for node assignment. High concavity points are characteristic of contours that enclose multiple objects and represent junctions where object intersection occurs. We leverage a concavity detection algorithm [9] in which concavity points are detected by computing the angle between vectors defined by sampling three consecutive points $(c_{w-1}, c_w, c_{w+1})$ on the contour. The degree of concavity/convexity is proportional to the angle $\theta(c_w)$, which can be computed from the dot product relation: $\theta(c_w) = \pi - \arccos\left(\frac{(c_w - c_{w-1}) \cdot (c_{w+1} - c_w)}{||(c_w - c_{w-1})|| \quad ||(c_{w+1} - c_w)||}\right)$. A point is considered to be

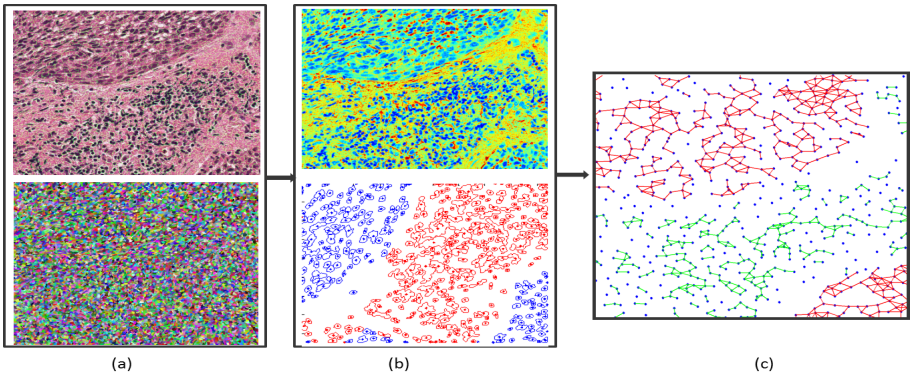


**Fig. 1.** (a) Original image with superpixel based partition of the entire image scene in the bottome panel, (b) (top panel) superpixels are classified as stromal and epithelial regions (cyan being epithelial, yellow being stroma and blue are nuclei), (bottome panel) nuclei clusters identified in the stroma and epithelial regions, and (c) establishing probabilistic links (edges) between identified nodes (third panel),where green edges are within stromal graphs and red edges are within epithelial graphs.

concavity point if $\theta(c_w) > \theta_t$, where $\theta_t$ is an empirically set threshold degree. Number of detected concavity points, $c_w \geq 1$, indicates presence of multiple overlapping/touching nuclei. In such cases, we consider the contour as one node, effectively identifying a cluster node. On each of the segmented cluster, the center of mass is calculated to represent the nuclear centroid.

### 2.3 Edge Connection

The last step is to build the links between the nodes that belong to the same group, i.e. epithelium or stroma, where the pairwise spatial relation between the nodes is translated to the edges (links) of SpACCl with a certain probability. The probability for a link between the nodes $u$ and $v$ reflects the Euclidean distance $d(u, v)$ between them and is given by

$$P(u, v) = d(u, v)^{-\alpha}, \tag{1}$$

where $\alpha$ is the exponent that controls the density of a graph. Probability of 2 nodes being connected is a decaying function of the relative distance. Since the probability of distant nuclei being connected is less, we can probabilistically define the edge set $E_i$ such that

$$E_i = \{(u, v) : r < d(u, v)^{-\alpha}, \forall u, v \in V_i\}, \tag{2}$$

where $r$ is a real number between 0 and 1 that is generated by a random number generator. In establishing the edges of SpACCl, we use a decaying probability function with an exponent of $-\alpha$ with $0 \leq \alpha$. The value of $\alpha$, set empirically (10-fold cross validation process), determines the density of the edges in the graph;

larger values of $\alpha$ produce sparser graphs. On the other hand as $\alpha$ approaches to 0 the graphs become densely connected and approach to a complete graph. Within each tissue region,$i$, there will be $i$ SpACCl graphs, which in turn are comprised of multiple sub-graphs as illustrated in Figure 1(c).

### 2.4 Subgraph Construction

SpaCCl's topological space decomposes into its connected components. The connectedness relation between two pairs of points satisfies transitivity, i.e., if $u \sim v$ and $v \sim w$ then $u \sim w$, which means that if there is a path from $u$ to $v$ and a path from $v$ to $w$, the two paths may be concatenated together to form a path from $u$ to $w$. Hence, being in the same component is an equivalence relation (defined on the vertices of the graph) and the equivalence classes are the connected components. In a nondirected graph $G_i$, a vertex $v$ is reachable from a vertex $u$ if there is a path from $u$ to $v$. The connected components of $G_i$ are then the largest induced subgraphs of $G_i$ that are each connected.

## 3 Feature Mining from SpACCl

Within each image, two separate graphs, $G^e$ and $G^s$ corresponding to the stromal and epithelial regions are obtained. We then extract both global and local graph metrics (features) from the subgraphs. These measurements are then averaged over the entire graph, $G^e$ and $G^s$ respectively. Table 1 summarizes the features we extract, along with their governing equations, and their histological significance and motivation.

## 4 Experimental Design and Results

### 4.1 Data Description

Using a tissue microarray cohort of 160 p16+ oSCC (punched twice with either 0.6 mm or 2 mm cores) with clinical follow up, digitally scanned H&E images at 400X magnification were annotated by an expert pathologist based on whether or not disease progression had occurred. There were 160 p16+ patients on the array – 141 cases of non-progressors and 19 progressors.

### 4.2 Classifier Training and Comparative Evaluation

We extract an identical set of features (listed in Table 1) from $G^e$ and $G^s$, $\mathbf{F} = \{F^e, F^s\}$ from $G_i$. The optimal feature set $\mathbf{Q}^{opt}$ was identified via mRMR feature selection scheme presented in [10], where *clustering D* and *average eccentricity* from $F^e$ and *number of central points* from $F^s$ were identified as the top most performing features. We then evaluated the accuracy of each of $G^s$ and

**Table 1.** Description of the features extracted from SpACCl. Identical features are extracted from each of $G^s$ and $G^e$ respectively.

| SpACCl Feature | Description | Relevance to Histology |
|---|---|---|
| Clustering Coeff C | Ratio of total number of edges among the neighbors of the node to the total number of edges that can exist among the neighbors of the node per node. $\tilde{C} = \frac{\sum_{u=1}^{\|V\|} C_u}{\|V\|}$, where $C_u = \frac{\|E_u\|}{\binom{k_u}{2}} = \frac{2\|E_u\|}{k_u(k_u-1)}$ | Nuclei Clustering |
| Clustering Coeff D | Ratio of total number of edges among the neighbors of the node and the node itself to the total number of edges that can exist among the neighbors of the node and the node itself per node. $\tilde{D} = \frac{\sum_{u=1}^{\|V\|} D_u}{\|V\|}$, where $D_u = \frac{k_u+\|E_u\|}{\binom{k_u+1}{2}} = \frac{2(k_u+\|E_u\|)}{k_u(k_u+1)}$ | |
| Giant Connected Comp | Ratio between the number of nodes in the largest connected component in the graph and the total number of nodes (Global) | |
| Average Eccentricity | $\frac{\sum_{u=1}^{V} \epsilon_u}{\|V\|}$ where eccentricity of the $u^{th}$ node $\epsilon_u$, $u = 1 \cdot \|V\|$, is the maximum value of the shortest path length from node $u$ to any other node in the graph. | Compactness of Nuclei |
| Percent of Isolated Points | Percentage of the isolated nodes in the graph, where an isolated node has a degree of 0 | |
| Number of Central Points | Number of nodes within the graph whose eccentricity is equal to the graph radius | |
| Skewness of Edge Lengths | Statistics of the edge length distribution in the graph | Spatial Uniformity |

$G^e$ in distinguishing progressors and non-progressors in oropharyngeal SCC by employing a SVM classifier trained on $\mathbf{Q}^{opt}$. We further trained SVM on $F^e$ and $F^s$ separately. A randomized 3 fold cross-validation involving 10 runs was used for mRMR feature selection and training the SVM. To demonstrate that class imbalance did not affect our classifier, we reported accuracy and positive predictive value (PPV) for the classifier for both the progressor and non-progressor cases.

To investigate the significance of encoding pairwise spatial relation between the nodes, we compared SpACCl against Voronoi and Delaunay graph based features (see [1]). Figure 2 illustrates tissue images of p16+ oropharyngeal progressor and non-progressor cases with associated of VT and SpACCl results. SpACCl provides a sparser and more localized representation of nuclear architecture compared to VT. SpACCl also prohibits the traversal of epithelium and stroma by constructing separate graphs for each region. The SVM classifier based of $\mathbf{F}^{opt}$ achieved a maximum accuracy of $90.2 \pm 1.2\%$ (Table 2). Our findings also suggest that the stromal SpaCCl graph features, $F^s$, can independently distinguish progressors and non-progressors in this population with an accuracy of 68% which suggests that stromal nuclear architecture is indeed informative in terms of predicting disease aggressiveness.

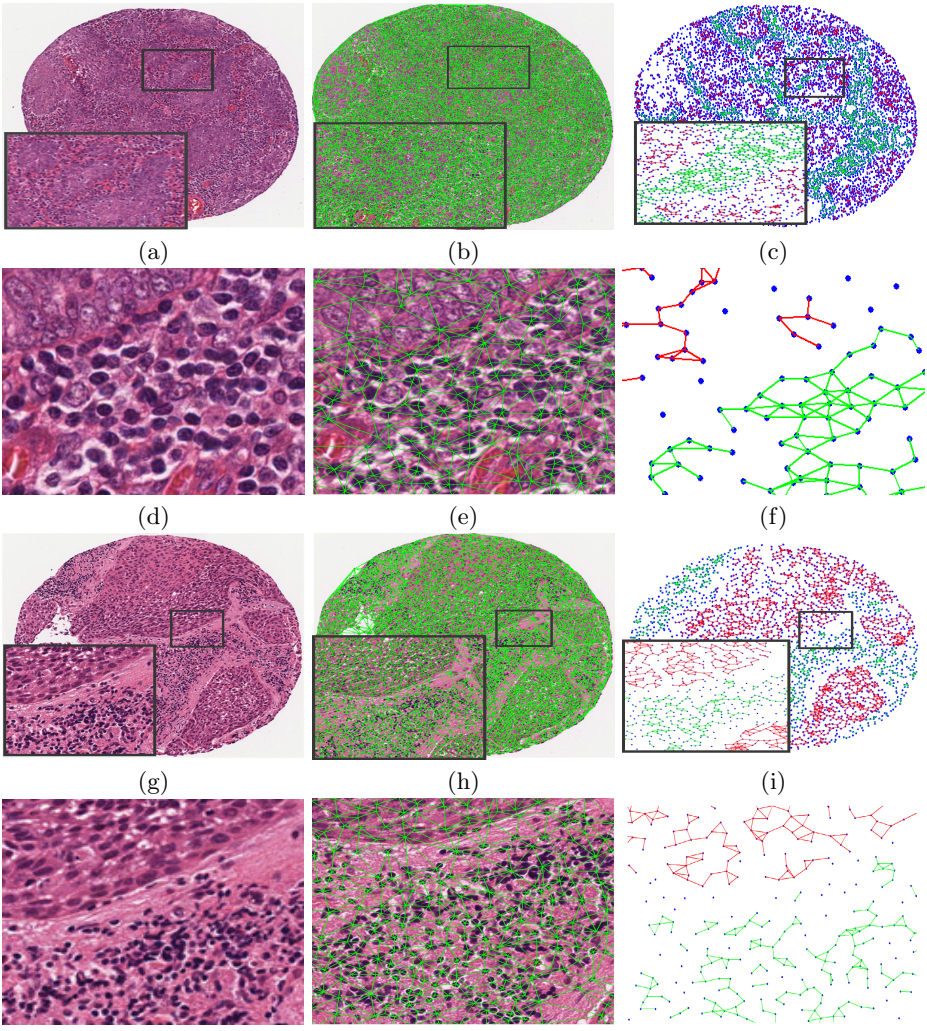


**Fig. 2.** Representative TMA core images of (a) progressor OSCC with enlarged ROIs in (d) and (g) Non-Progressor image. (b) and (h) represent the Delauny graphs of (a) and (g) with enlarged ROIs in (e) and (k). SpACCl graphs for (a) and (g) are shown in (c) and (i) with enlarged ROIs in (f), (l) respectively.

**Table 2.** Accuracies compared against the Voronoi and Delauny graphs, and PPV values for Progressors (p) and Non-Progressor (NP)

| | Tradiational Graphs | | SpaCCl | | |
|---|---|---|---|---|---|
| | Voronoi | Delauny | $F^e$ | $F^s$ | $Q^{opt}$ |
| Accuracy | $74.4 \pm 0.6\%$ | $76.7 \pm 0.7\%$ | $86.2 \pm 1.2\%$ | $68.1 \pm 0.2$ | $\mathbf{90.1 \pm 1.5}$ |
| PPV (P) | $77.9 \pm 1.6\%$ | $74.3 \pm 0.5\%$ | $85.2 \pm 1.6\%$ | $76.1 \pm 0.2$ | $89.4 \pm 0.2$ |
| PPV (NP) | $76.9 \pm 0.9\%$ | $76.3 \pm 1.5\%$ | $82.5 \pm 2.6\%$ | $78.7 \pm 0.2$ | $86.5 \pm 0.5$ |

## 5 Concluding Remarks

In this work we presented a new sptially-aware Cell Cluster (SpACCl) graph algorithm and showed its application in the context of predicting disease aggressiveness in p16+ oropharyngeal cancers. SpACCl allows for implicit construction of two separate nuclear graphs within each of the stromal and epithelial regions, which enabled us to independently explore contribution of nuclear architecture within the two regions. It also allowed us to explore the combined contributions of stromal and epithelial nuclear architecture in predicting disease aggressiveness. The results obtained in this work with SpACCl suggest that (a) stromal nuclear architecture has a role to play in predicting patient outcome for this class of tumors and that (b) combining stromal and epithelial nuclear architecture results in better prognostic accuracy compared to graph measures obtained from either the stromal or epithelial graphs alone.

## References

1. Doyle, S., et al.: Cascaded discrimination of normal, abnormal, and confounder classes in histopathology: Gleason grading of prostate cancer. BMC Bioinformatics, 1284–1287 (2012)
2. Epstein, J., et al.: The 2005 international society of urological pathology (isup) consensus conference on gleason grading of prostatic carcinoma. American J. Surgical Path. 29(9), 1228–1242 (2005)
3. Tabesh, A., et al.: Multifeature prostate cancer diagnosis and gleason grading of histological images. IEEE TMI 26(10), 1366–1378 (2007)
4. Lewis, et al.: Tumor cell anaplasia and mutinucleation are predictors of disease recurrence in oropharyngeal cancers. Am. J. Path. (2012)
5. Gunduz, et al.: The cell graphs of cancer. Bioinformatics 20, i145–i155 (2004)
6. Liu, E.T., et al.: N. Engl. J. Med. 357, 2537–2538 (2007)
7. Beck, et al.: Systematic analysis of Breast Cancer morphology uncovers stromal features associated with survival. Sci. Transl. Med. (2011)
8. Ren, X., et al.: Learning a classification model for segmentation. In: ICCV, vol. 1, pp. 10–17 (2003)
9. Fatakdawala, H., et al.: Expectation Maximization driven Geodesic Active Contour with Overlap Resolution (EMaGACOR). IEEE TBME 57(7), 1676–1689 (2010)
10. Peng, H., et al.: Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. PAMI 27, 1226–1238 (2005)