Using Manifold Learning for Content-Based Image Retrieval of Prostate Histopathology

Scott Doyle\textsuperscript{1}, Mark Hwang\textsuperscript{1}, Shivang Naik\textsuperscript{1}, Michael Feldman\textsuperscript{2}, John Tomaszeweski\textsuperscript{2}, and Anant Madabhushi\textsuperscript{1}

\textsuperscript{1} Department of Biomedical Engineering, Rutgers University, USA, *\textsuperscript{2} Department of Surgical Pathology, University of Pennsylvania, USA.

Abstract. We present a content-based image retrieval (CBIR) system for the retrieval of digitized images of prostate histopathology. In our study we extract nearly 600 content-based features from digitized images of prostate histology that describe the architecture, morphology, and texture of the images. Manifold learning is used to map the high dimensional data from a non-linear manifold onto a low-dimensional linear subspace, in which object adjacencies are preserved. We analyze a set of 56 images including 19 benign epithelium, 23 Gleason grade 3, and 14 Gleason grade 4 images. We use feature subsets of 483 image texture features, 44 gland morphology features, and 49 graph-based features to determine if some feature sub-groups are more discriminating than others. The 576-dimensional feature space was projected into different subspaces of dimensions ranging from 1-10 using 7 different manifold learning algorithms. The idea behind using these different manifold learning algorithms was to identify the optimal lower-dimensional data reconstruction within which to evaluate object similarity. The system was tested by selecting a query image from the dataset and evaluating its similarity to the remaining images using a linear metric in the reduced space. Mean average retrieval precision of 0.573 was obtained for Gleason grade 3, 0.418 for Gleason grade 4, and 0.566 for benign epithelium using morphological features, which perform statistically significantly better than other feature subsets. Highest precision was obtained using principal component analysis and laplacian eigenmaps.

1 Introduction

Prostate cancer is the most commonly diagnosed cancer among males in the U.S., with 200,000 new cases and 27,000 deaths predicted for 2007 (source: American Cancer Society). Manual examination of prostate biopsy samples by an expert pathologist is the current gold standard of prostate cancer diagnosis. In the U.S., the most common system of grading prostate tissue (assessing degree of

\* This work was supported by grants from the Coulter foundation, Busch Biomedical Award, Cancer Institute of New Jersey, New Jersey Commission on Cancer Research, and the National Cancer Institute (R21CA127186-01, R03CA128081-01).

Corresponding author: scottdo@eden.rutgers.edu
Fig. 1. Examples of (a) Gleason grade 3 tissue, (b) Gleason grade 4 tissue, (c) a gland from (a) magnified, (d) a gland from (b) magnified, (e) a benign gland, and (f) an illustration of the lumen and nuclei comprising the gland in (e).

malignancy) is the Gleason scale [1], which assigns grades from 1 (relatively benign tissue) to 5 (highly invasive cancer).

The Gleason paradigm illustrates how cancer differs in terms of architecture (spatial arrangement of nuclei and glands within the tissue with respect to their centers of mass) and morphology (shape and size of glands and nuclei) as malignancy progresses. Glands and nuclei both express architectural and morphological changes in the development of cancer [2]. An example of tissue regions of Gleason grade 3 tissue is shown in Fig. 1 (a), grade 4 tissue in Fig. 1 (b), a single grade 3 gland in Fig. 1 (c), and a grade 4 gland in Fig. 1 (d). A gland from benign epithelial tissue is shown in Fig. 1 (e). An illustration of the lumen (white region) and nuclei (grey ellipses) of the gland in Fig. 1 (e) is shown in (f).

A number of studies have identified issues with the Gleason system, including high degrees of observer variability, with tissue under-grading as high as 48% [3]. Because of the diagnostic importance of Gleason grading, a quantitative system for assisting pathologists in analyzing histopathology will improve patient care by providing an accurate and standardized grading tool.

A great deal of research has focused on creating content-based image retrieval (CBIR) systems to analyze medical image data [4]. Manifold learning (ML) methods are commonly used to map the high dimensional image feature data into a low-dimensional space [4]. These methods project the data from a high to a low dimensional space based on similarity metrics, which are dependent upon assumptions about the data. For example, linear methods assume that the underlying manifold of the data is linear, while non-linear methods assume non-linearity of the data. Many ML algorithms have been developed over the years, and each uses a different method to calculate object adjacencies. Most CBIR systems employ principal component analysis [4].

Computer-aided diagnostic (CAD) tools have recently been developed to quantify prostate image data from high-resolution ex vivo MRI [5]. In [6], we presented a CAD system to determine the Gleason grade of cancerous areas on prostate histology using content-based image features including architectural, morphological and textural features. In this work, we present a CBIR system for prostate histopathology based on those features. We evaluate the performance of the system by looking at nearly 600 features which characterize the texture,
Fig. 2. Overview and organization of our CBIR system for automated retrieval of prostate histopathology images.

morphology, and architecture of histopathological images. 7 manifold learning methods are used to reduce the data to between 1 and 10 different dimensions. The system is tested on 56 studies labeled as Gleason grade 3 (23 studies), grade 4 (14 studies), and benign epithelium (19 studies) by an expert pathologist. The main contributions of this work are:

- A CBIR system for prostate histopathology that employs manifold learning to reduce a set of images from a high-dimensional, non-linear feature space to a low-dimensional subspace;
- A novel set of content-based image features that capture characteristics defined in the Gleason scheme (such as morphology) as well as those not analyzed in clinical pathology (such as texture and graph-based features); and
- Investigation into the effect of manifold learning algorithm choice and dimensionality on the ability of a CBIR system to retrieve relevant images.

An overview of the system is described in Section 2. In Section 3 we describe our feature extraction methods. The manifold learning algorithm is described in Section 4. Results are given in Section 5 and concluding remarks in Section 6.

2 System Overview

An overview of our system is shown in Figure 2. Offline, a database of histopathological prostate images is constructed by extracting graph-based, textural, and morphological features from a series of images comprising a set of tissue studies that have been graded by an expert pathologist. These images are then projected into a low-dimensional space using one of several manifold learning (ML) methods to evaluate object adjacency. In this study, we consider principal component analysis (PCA), multidimensional scaling (MDS), graph embedding (GE), Isomaps (ISO), local linear embedding (LLE), kernel-based PCA (k-PCA), and laplacian eigenmaps (LE). Online, a query image is run through the feature extraction algorithm and plotted into the reduced dimensional space using the same ML algorithm that was used in the building of the database. If the ML
method has accurately estimated object adjacency, the query image should be mapped close to database images from the same class. Finally, in the low dimensional space, a linear Euclidean distance metric is used to rank the database images in order of similarity to the query image. A retrieved image is “relevant” if it is the same class as the query image (Gleason grade 3, grade 4, or benign epithelium), and “irrelevant” otherwise. We calculate mean average precision (MAP) values to determine which ML algorithm yields the best retrieval precision, as well as which feature classes perform best in describing the images in the CBIR system. A Student’s t-test is used to determine if different feature subsets produce statistically significant differences in performance.

3 Feature Extraction

Hematoxylin and eosin stained prostate biopsy cores are imaged on a high resolution whole slide digital scanner at 40x magnification and saved on a computer workstation. We denote a tissue region $R$ by a digital image $C^R = (C, f)$ where $C$ is a 2D grid of image pixels $c \in C$ and $f$ is a function that assigns an intensity to $c$. $C^R$ comprises $k$ glands with centroids at pixels $c_1^R, c_2^R, \cdots, c_k^R$. $C^R$ also comprises $m$ nuclei (grey ellipsoids in Fig. 1 (f)) with centroids at manually labeled pixels $c_1^m, c_2^m, \cdots, c_m^m$.

3.1 Graph-based Features to Describe Nuclear Architecture

A. Nuclear Features We compute the following 25 features directly from the spatial location of the centroids of the nuclei in $C^R$ to characterize nuclear proliferation. (1) The density of the nuclei in $C^R$ is computed as $D = \frac{m}{|C|}$, where $|C|$ is the cardinality of $C$. (2) We denote by $S_K$ the set of $K$-nearest neighbors of nuclear centroid $c_a^m$ where $K \in \{3, 5, 7\}$ and $a \in \{1, 2, \cdots, m\}$. Average nuclear distance of $c_a^m$ is given by $d_{c_a^m, K} = \frac{1}{|S_K|} \sum_{c \in S_K} ||c_a^m - c||$. The overall average nuclear distance $\mu_{c_a^m, K}^d = \frac{1}{m} \sum_{a \in \{1, 2, \cdots, m\}} d_{c_a^m, K}$ and standard deviation $\delta_{c_a^m, K}^d$ over all $a \in \{1, 2, \cdots, m\}$ is calculated. In addition, a measurement of disorder
quantifying the variation of $d_{e,n,K}$ for all $a$ is given as $\Psi_{n,K}^d = 1 - \frac{1}{1 + \frac{\sigma_n^2}{\mu_n^2}}$), giving an additional 9 features for $C^R$. We denote by $B_{c_n^a,r}$ a ball of pixels with radius $r$ centered on $c_n^a$. The number of pixels corresponding to nuclear centroids $c_n^a, j \neq a, j \in \{1, 2, \cdots, m\}$ in $B_{c_n^a,r}$ are counted and the sum denoted as $Q_{c_n^a,r}$. The mean and standard deviation of $Q_{c_n^a,r}$ for $a \in \{1, 2, \cdots, m\}$ are denoted by $\mu_{n,r}^Q$ and $\sigma_{n,r}^Q$. The measurement of disorder $\Psi_{n,r}^Q$ is also calculated as described above for $\Psi_{n,K}^d$. In this study, we use values of $r \in \{10, 20, \cdots, 50\}$ which were determined empirically.

B. Voronoi Diagram The Voronoi diagram $V$ partitions $C^R$ with a series of polygons. Polygon $P_{c_n^a}$ is constructed around $c_n^a$, creating a tesselation of $C^R$. Every pixel is assigned to a polygon and every polygon is associated with a nuclear centroid. Each $P_{c_n^a}$ has $e$ unique edges $E_{b+1}^V, E_{b+2}^V, \cdots, E_{b+e}^V$ between all adjacent vertices with corresponding edge lengths $l_{b+1}^V, l_{b+2}^V, \cdots, l_{b+e}^V$ and chord lengths $H_1, H_2, \cdots, H_h$ between all nonadjacent vertices. Each $P_{c_n^a}$ has a perimeter $l_P^V = \sum_{i=1}^e l_i^V$, total chord length $l_H^V = \sum_{i=1}^h H_i$, and total area $A^V = |P_{c_n^a}|$. We compute the average, standard deviation, ratio of minimum to maximum value, and disorder for $A^V$, $l_P^V$, and $l_H^V$ of each $P_{c_n^a}$ in $C^R$, giving 12 features.

C. Delaunay Triangulation The Delaunay graph $D$ is a graph constructed so that any two unique nuclear centroids $c_n^a$ and $c_n^b$, where $a, b \in \{1, 2, \cdots, m\}$, are connected by an edge $E_{a,b}^D$ if their associated polygons in $V$ share a side. The average, standard deviation, minimum to maximum ratio, and disorder of the areas and edge lengths are computed for all triangles in $D$, giving 8 features.

D. Minimum Spanning Tree A spanning tree $S$ of $D$ is a subgraph which connects all $c_n^a, a \in \{1, 2, \cdots, m\}$ together. A single $D$ can have many $S$. The minimum spanning tree denoted by $S^T$ has a total length less than or equal to the total length of every other spanning tree. This graph describes the spacing and distance between nuclei, which is expected to change as cancer progresses and nuclei proliferate. We compute the average, standard deviation, minimum to maximum ratio, and disorder of the edge lengths in $S^T$ to obtain an additional 4 and a total of 24 graph-based features.

3.2 Gland Architecture and Morphology

A. Co-Adjacency Features We denote as $c_1^g, c_2^g, \cdots, c_k^g$ the centroids of $k$ glands within $C^R$, and construct a co-adjacency matrix $W$ wherein the value of row $u$, column $v$, $W(u, v) = ||c_u^g - c_v^g||$, $u, v \in \{1, 2, \cdots, k\}$, and $W \in \mathbb{R}^{k \times k}$ where $||\cdot||$ denotes Euclidean distance. $W$ describes the inter-gland spatial relationships in a manner similar to the co-occurrence matrix proposed by Haralick to describe spatial relationships between pixel intensity values. 13 of Haralick’s second-order features are calculated from $W$: angular second moment, contrast, correlation,
variance, entropy, sum average, sum variance, sum entropy, difference variance, difference entropy, difference moment, and two measurements of correlation.

B. Morphological Features The lumen area is surrounded by a boundary $B$ obtained via a level-set algorithm, where the initial contour is initialized by the user inside the gland near the lumen area (the white region in Figs. 4 (a) and (b)) and is allowed to evolve to its final position (black line in Figs. 4 (a) and (b)). The lumen area is We denote as $l_B$ the length of the gland boundary $B$. The distance from the centroid of the gland $c_g$ to boundary pixel $c_B^\alpha$ is denoted $d(c_g, c_B^\alpha)$, where $c_B^\alpha \in B$. We compute the average and maximum of $d(c_g, c_B^\alpha)$ over $\alpha \in \{1, 2, \cdots, \beta\}$. We also obtained the fractal dimension of the gland boundary. We picked intermediate points $c_B^\gamma \in B$ where $\gamma \in \{3, 6, 9\}$ on $B$ and linearly interpolated between these points to obtain length $l_B^\gamma$. The fractal dimensions are obtained as $l_B^\gamma / l_B^\alpha$. The following values are calculated for each gland in $C^R$: gland area $A_G$, lumen area $A_L$, boundary length $l_B$, number of nuclei surrounding the lumen, and number of layers of nuclei encircling the gland. A number of other features are obtained by considering ratios and combinations of $A_G$, $A_L$, $l_B$, and $d(c_g, c_B^\alpha)$ for $\alpha \in \{1, 2, \cdots, \beta\}$, generating 8 values. The average, standard deviation, and measurement of disorder of these 8 features for all k glands is calculated as described in Section 2.1 to yield 24 features for $C^R$. We also calculate the standard deviation and variance of $d(c_g, c_B^\alpha)$ over $\alpha \in \{1, 2, \cdots, \beta\}$, $l_B/l_B^\gamma$, and $(l_B^\gamma)^2/A_G$ for each gland. Finally, for any point on the boundary $c_B^\alpha \in B$ and its adjacent points $c_B^{\alpha-1}$ and $c_B^{\alpha+1}$, the smoothness factor is calculated as $U_\alpha = d(c_g, c_B^\alpha) - (d(c_g, c_B^{\alpha-1}) + d(c_g, c_B^{\alpha+1})) / 2$. The sum of $U_\alpha$ for $\alpha \in \{1, 2, \cdots, \beta\}$ is calculated for each gland. The average of these 7 values are computed for k glands in $C^R$ giving 7 morphological features.

Fig. 4. Glands from (a) Gleason grade 3 tissue and (b) Gleason grade 4 tissue, with lumen boundaries shown in white and nuclear boundaries in black.

3.3 Texture Descriptors

The average, median, standard deviation, and the range of $f(c)$ is computed for all $c \in C$ for each of the three image color channels, yielding 12 first-order statistical features. A co-occurrence matrix $Z \in \mathbb{R}^{M \times M}$ is constructed for $C^R$ where $M \times c \in C$ is the maximum pixel value of $C$. The value in $Z(f(b), f(c))$ where $b, c \in C$ is given by the number of times intensities $f(b)$ and $f(c)$ occur within a fixed displacement of 1 pixel from each other at any orientation. From $Z$ we extract 39 Haralick features from each image. Finally, a family of 2D Gabor filter kernels is created from a modulating function, which is constructed from a Gaussian function modulated by a sinusoid.
Table 1 shows the different feature subsets used in this study. Graph-based features are extracted from the Voronoi Diagram \( V \), Delaunay Triangulation \( D \), and Minimum Spanning Tree \( S^T \). Gland architecture and morphological features are extracted from the co-adjacency matrix \( W \) and the boundary \( B \). Texture features include statistical, Haralick, and Gabor features.

<table>
<thead>
<tr>
<th>Feature Subset</th>
<th>Number of Features</th>
<th>Major Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Graph-Based</td>
<td>49</td>
<td>( V, D, S^T )</td>
</tr>
<tr>
<td>Gland Architecture &amp; Morphology</td>
<td>44</td>
<td>( W, B )</td>
</tr>
<tr>
<td>Texture Descriptors</td>
<td>483</td>
<td>Statistical, Haralick, Gabor</td>
</tr>
</tbody>
</table>

Table 1. The different feature subsets used in this study, along with the major features within each feature type.

4 Manifold Learning

Manifold learning (ML) methods map a set of data from \( N \) dimensions to \( M \) dimensions, where \( M << N \), preserving the high-dimensional relationships between data points. In this study, we employ two groups of ML methods:

**Linear Methods** are used to calculate object adjacency for data that lie on a linear high-dimensional manifold. In this study, we employ PCA (commonly employed in CBIR systems [4]) and MDS to linearly project the data.

**Non-linear Methods** are used to accurately compare object similarities that lie on a non-linear manifold. Recent studies have found that for genome expression data, non-linear methods outperform linear methods in mapping out the true class relationships [7]. In this study, we employ GE, ISO, LLE, k-PCA, and LE non-linear ML methods.

We also analyze the full unreduced data set to evaluate whether retrieval in the reduced dimensional space is improved over the unreduced space.

5 Results

**Precision and Recall** By iterating through all of the returned images, we evaluate the system using precision vs. recall (PR) graphs [4], where precision is the ratio of the number of relevant images retrieved to the total number of retrieved images and recall is the ratio of the number of relevant images to the total number of relevant images in the database. A recall of 1.0 is obtained when all images are retrieved from the database, while a precision of 1.0 is obtained if all retrieved images are relevant. The retrieved images are sorted in order of increasing Euclidean distance from the query image, so that the first
image returned is most similar to the query. Each image is queried against the remaining images in the database, and we iterate through each of the returned images to generate a PR graph. The PR graphs obtained for all images of the same class are averaged together to generate the final PR curve for that class.

**Mean Average Precision** We also calculate the *Mean Average Precision* (MAP), an average of the precision for all returned images normalized by the total number of relevant images in the database. MAP values are shown in Table 2. Each row in the table represents the MAP obtained using a particular feature set and class of the query image, and each column shows the ML method used. For each class, the highest MAP values are shown in boldface. In all three classes, the highest MAP values were obtained when using only morphological features. PCA yielded the highest MAP for Gleason grades 3 and 4, while LE produced the highest MAP for benign epithelium. MDS performed as well as PCA when Gleason grade 3 was the query image, but performance decreased when Gleason grade 4 was the query image. Table 3 shows the results from a two-tailed paired Student’s t-test comparing MAP values obtained using morphological features alone to those of the other feature subsets. Shown are the results from two of the ML methods analyzed, MDS and GE. In most cases, the values indicate that morphological features result in a statistically significant change in MAP values.

**Qualitative Results** Results from manifold learning are shown in Figure 5. Feature vectors are plotted in the 3-dimensional subspace obtained through (a) MDS and (c) PCA, as applied to morphological features, which performed the best in quantitative analysis. Points in the scatter plot correspond to Gleason grade 3 (green circles), Gleason grade 4 (blue squares), and benign epithelium (red triangles). The boxed region contains the majority of points and is shown
<table>
<thead>
<tr>
<th>Query Image</th>
<th>MDS</th>
<th>GE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Texture</td>
<td>Graph</td>
</tr>
<tr>
<td>Gleason Grade 3</td>
<td>6.82E-10</td>
<td>1.63E-09</td>
</tr>
<tr>
<td>Gleason Grade 4</td>
<td>5.68E-05</td>
<td>6.64E-04</td>
</tr>
<tr>
<td>Benign Epithelium</td>
<td>4.01E-03</td>
<td>4.70E-02</td>
</tr>
</tbody>
</table>

Table 3. Results of a two-tailed paired Student’s t-test, comparing MAP values for morphology against different subsets of features using two different ML methods. P-values less than 0.05 indicate significantly different results.

![Image](image-url)

**Fig. 5.** Scatter plots obtained through (a) MDS and (c) PCA, with a closeup of the boxed region. The PR curve for all classes obtained using (b) MDS and (d) PCA. Shown are images from Gleason grade 3 (green circles), Gleason grade 4 (blue squares), and benign epithelium (red triangles). Class clusters are manually indicated in black.

magnified. Class clusters show separation between the classes when using the reduced feature vectors. Because of their similar appearance, images representing Gleason grades 3 and 4 generally appear very close to one another in the reduced space. Also shown are PR curves obtained using (Fig. 5 (b)) MDS and (Fig. 5 (d)) PCA, using morphological features. For each of the ML methods, the precision of benign epithelium is high for a low number of returned images and decreases as more images are returned, while the Gleason grade 3 images and grade 4 images have consistent precision as the number of retrieved images increases.
6 Concluding Remarks

In this paper we presented a CBIR system to retrieve images from a database using a novel set of nearly 600 features and 7 different manifold learning methods. In this work, we found:

- Morphological features yield the highest MAP for all classes of query images;
- MAP was highest using a subspace obtained using PCA for Gleason grades 3 and 4 and LE for benign epithelium; and
- ML methods improve retrieval precision over the un-reduced feature space.

Note that unlike previous approaches that have sought to distinguish between low- and high-grade cancers we have focused on the clinically significant problem which is also the reason for inter- and intra-observer grading variability among pathologists, namely distinguishing grades 3 and 4. We are currently developing an algorithm for the automated detection and segmentation of gland regions in prostate histological images, which will fully automate the feature extraction process. This algorithm will integrate low, high, and domain-level knowledge to automatically detect gland regions, initialize a level-set curve, and evolve the curve to the interior nuclear boundary.

Manifold learning normally requires re-computation of the low-dimensional space when a sample is added to the dataset (i.e. when comparing a query image to the database). Law and Jain [8] have proposed a method of incremental dimensionality reduction which does not require a re-computation of the embedding, which would allow query images to be quickly compared with the database images in the low dimensional space. Finally, a comprehensive study of optimal feature selection will be possible as more data becomes available for analysis.

References