Gleason Grading of Prostate Histology Utilizing Manifold Regularization via Statistical Shape Model of Manifolds

Rachel Sparks and Anant Madabhushi
rsparks@eden.rutgers.edu, anantm@rci.rutgers.edu
Department of Biomedical Engineering, Rutgers University, USA

1. ABSTRACT

Gleason patterns of prostate cancer histopathology, characterized primarily by morphological and architectural attributes of histological structures (glands and nuclei), have been found to be highly correlated with disease aggressiveness and patient outcome. Gleason patterns 4 and 5 are highly correlated with more aggressive disease and poorer patient outcome, while Gleason patterns 1-3 tend to reflect more favorable patient outcome. Because Gleason grading is done manually by a pathologist visually examining glass (or digital) slides, subtle morphologic and architectural differences of histological attributes may result in grading errors and hence cause high inter-observer variability. Recently some researchers have proposed computerized decision support systems to automatically grade Gleason patterns by using features pertaining to nuclear architecture, gland morphology, as well as tissue texture. Automated characterization of gland morphology has been shown to distinguish between intermediate Gleason patterns 3 and 4 with high accuracy. Manifold learning (ML) schemes attempt to generate a low dimensional manifold representation of a higher dimensional feature space while simultaneously preserving nonlinear relationships between object instances. Classification can then be performed in the low dimensional space with high accuracy. However ML is sensitive to the samples contained in the dataset; changes in the dataset may alter the manifold structure. In this paper we present a manifold regularization technique to constrain the low dimensional manifold to a specific range of possible manifold shapes, the range being determined via a statistical shape model of manifolds (SSMM). In this work we demonstrate applications of the SSMM in (1) identifying samples on the manifold which contain noise, defined as those samples which deviate from the SSMM, and (2) accurate out-of-sample extrapolation (OSE) of newly acquired samples onto a manifold constrained by the SSMM. We demonstrate these applications of the SSMM in the context of distinguishing between Gleason patterns 3 and 4 using glandular morphologic features in a prostate histopathology dataset of 58 patient studies. Identifying and eliminating noisy samples from the manifold via the SSMM results in a statistically significant improvement in classification accuracy (CA), 93.0 ± 1.0% with removal of noisy samples compared to a CA of 90.9 ± 1.1% without removal of samples. The use of the SSMM for OSE of new independent test instances also shows statistically significant improvement in CA, 87.1 ± 0.8% with the SSMM compared to 85.6 ± 0.1% without the SSMM. Similar improvements were observed for the synthetic Swiss Roll and Helix datasets.

2. INTRODUCTION

Blinded needle sextant biopsy is the current gold standard for prostate cancer (CaP) diagnosis; each biopsy yields 12-18 needle cores which are then analyzed under a microscope by a pathologist.\textsuperscript{1,2} If CaP is identified, a pathologist will then assign a Gleason score to the biopsy samples, with higher scores corresponding to more aggressive CaP.\textsuperscript{3} Gleason score is determined as a summation of the two most prevalent Gleason patterns, hence Gleason score has a range of 2-10 where scores 2-4 represent low grade CaP, scores 5-6 represent intermediate grade CaP, and 7-10 represent high grade CaP.\textsuperscript{4} Gleason score aids in determining the course of treatment, patients with less aggressive CaP (Gleason score 6) may be enrolled in active surveillance programs while patients with more aggressive CaP (Gleason score 7 and above) will undergo radiation therapy or surgery.\textsuperscript{5}

Low Gleason patterns are characterized by a coherent spatial architecture with distinct gland lumen surrounded by cell nuclei.\textsuperscript{3,6} For higher Gleason patterns, the arrangement and morphology of histological structures begins to breakdown with gland lumen becoming indistinct and crowded with an increase in the concentration of cell nuclei. Manually distinguishing intermediate Gleason patterns 3 and 4 accurately is a difficult problem; previous studies have reported an inter-observer agreement between pathologists of 0.47-0.64 (reflecting low to moderate agreement).\textsuperscript{7,8}
Figure 1. (a) 3D Swiss Roll dataset with Gaussian noise added to 2% of samples in the dataset. (b) 2D manifold \( M \) in the absence of noise. This manifold structure best preserves the relationships between samples in the original high dimensional space. (c) Manifold \( \hat{M} \) found by applying ML to a dataset containing noise and (d) the manifold \( \tilde{M} \) found by regularization of \( M \) via the SSMM.

Availability of digital prostate histology samples\(^9\) has lead to the development of computer assisted decision support tools to quantify subtle changes in prostate tissue and thus accurately distinguish between Gleason patterns.\(^10–19\) Previous work which has quantified changes in prostate histology has utilized image texture,\(^10, 14, 18\) arrangement and morphology of nuclei,\(^12, 13, 15, 19\) or morphology of glands.\(^16, 17\) However, a large number of features are typically necessary to accurately determine Gleason grade for histopathology images, resulting in a high dimensional feature space.

Dimensionality reduction offers a way to overcome the curse of dimensionality by constructing a low dimensional space in which to perform classification while not compromising object-class relationships. Manifold learning (ML) refers to a class of nonlinear dimensionality reduction methods that aim to learn a low dimensional space in which to perform classification while not compromising object-class relationships. Manifold learning schemes tend to be sensitive to the dataset considered, and changes in the dataset may alter the structure of the learned manifold.\(^27\) Consider a sample \( o_i \in \mathbf{O} \) perturbed by some error \( \epsilon \); the new location for \( o_i \) would be \( \tilde{o}_i = o_i + \epsilon \). \( A \) would have to be altered such that \( \tilde{A}(\tilde{o}_i, o_j) = \psi(\tilde{o}_i, o_j) \) for all \( o_j \) contained in \( \mathbf{O} \), resulting in changes to \( 2N \) elements in \( A \). The manifold \( \tilde{M} \) learned from \( \tilde{A} \) will reflect those changes. Hence even a small change in \( \mathbf{O} \) may cause large changes to \( \tilde{M} \). Figure 2 demonstrates this phenomenon for a prostate histology dataset comprising 888 glands. Two manifolds were generated by applying ML to 90% of samples in the dataset (800 glands) such that for each manifold a different set of 88 samples were excluded. Each manifold has a...
distinct structure evident by (a) changes in the planar structure of the manifold and (b) changes in object-class relationships on the manifold, displayed as color differences between manifolds.

Consider a large dataset \( O \) from which the manifold \( M \) is generated. In the absence of knowing the true manifold, \( M \) is the best manifold representation to capture the relationships between samples in the dataset. If we consider a subset \( \hat{O} \subset O \) then \( \hat{O} \) can be used to create an alternative manifold \( \hat{M} \) which approximates \( M \). Manifold regularization constrains the structure of \( \hat{M} \) giving a better approximation of \( M \) and hence resulting in a better representation of the relationships between samples in \( O \).

In this work we present a statistical shape model of manifolds (SSMMs) to perform manifold regularization. The concept is that shape variation for a set of manifolds can be modeled with a statistical shape model (SSM). SSMs have been proposed to model shape variation in anatomical structures. In much the same way, we utilize a SSM to model which manifold shapes are statistically most likely to occur. The SSMM describes the mean shape and primary modes of variation for a series of different manifolds constructed by randomly sampling different parts of the same dataset. For a new, related dataset, the resulting manifold can be constrained to only the range of shapes dictated by the SSMM. Hence every sample on the new manifold is spatially constrained.

The remainder of the paper is organized as follows. Section 3 describes previous work in Gleason grading of prostate histology and manifold regularization. An overview of SSMM construction and its novel contributions are discussed in Section 4. Section 5 provides a review of key concepts in ML theory. Section 6 describes the methodology to construct a SSMM and its application to (a) identification of samples which contain noise and (b) out-of-sample extrapolation (OSE) of newly acquired samples onto the SSMM. Section 7 describes the experimental design and results for two synthetic datasets as well as a prostate histology dataset. Concluding remarks are presented in Section 8.
3. PREVIOUS WORK

3.1 Automated Gleason Grading

Pathologists perform Gleason grading of prostate cancer tissue specimens via qualitative, visual evaluation of a tissue section previously stained with Hematoxylin and Eosin (H& E). The primary discriminating traits of Gleason patterns on histopathology are the arrangement and morphology of the nuclei and glands within a tissue sample. In designing automated pattern recognition methods for distinguishing different Gleason patterns on histopathology, the key questions to consider are (1) what is the best feature set to distinguish between Gleason patterns? and (2) what is the best method to reduce the dimensionality of the feature set prior to classification?

Jafari et. al. characterized tissue patch texture via wavelet features and classified Gleason patterns with an accuracy of 97% for the best performing feature. Huang et. al. characterized tissue patch texture via Fractal Dimension and achieved an accuracy of 95%. However, a limitation of these approaches were that the image patches were manually selected to obtain regions which contained only one tissue class on the digitized slide. DiFranco et. al. characterized tissue patch texture for each color channel independently showing 90% accuracy classifying images on a per tile. Although tiles were automatically determined, tiles which contained more than one tissue class were removed from the dataset.

Structural features (as opposed to texture features) have also been explored by some researchers for automated categorization of Gleason patterns. Veltri et. al. and Ali et. al. showed that the quantitative characterization of the shape of individual nuclei on tissue microarrays can distinguish between Gleason patterns with high accuracy. In a preliminary study by Veltri et. al. characterization of manually segmented nuclei were able to distinguish between Gleason pattern 3, 4, and 5 with 73 – 80% accuracy. Ali et. al. automated the nuclear segmentation and classification giving 84% accuracy on 80 tissue microarrays. Doyle et. al. characterized manually selected image patches according to nuclear arrangement, reporting a predictive positive value of 76.0% in distinguishing between Gleason patterns 3, 4, and 5 within a multi-classification scheme. In previous work we have shown that gland morphology, quantified by Diffeomorphic Based Similarity (DBS), is able to distinguish between Gleason 3 and 4 patterns with 88% accuracy. DBS is calculated by constructing shape models for each gland contained in a set of histology images and then quantifying the differences between shape models. Tabesh et. al. combined gland morphology, texture features, color channel variance, and nuclear arrangement to classify different Gleason patterns with 81.0% accuracy. Golugula et. al. used proteomic data in conjunction with histology derived image features to distinguish between prostate cancer patients who following radical prostatectomy had biochemical recurrence within 5 years from those who did not.

Most automated Gleason grading systems are described by a high dimensional feature space. To perform accurate classification the high dimensional feature space must be reduced to a lower dimensional space. One approach to reduce the high dimensional feature space is to perform feature selection, thereby determining a small subset of the original feature space in which accurate classification can be performed. Diffranco et. al. utilized a random forest feature selection algorithm. Doyle et. al. utilized a cascaded classification approach to perform feature selection for a series of pairwise classification tasks. A limitation of these approaches is that features which are excluded may contain important classification information, hence their removal can diminish classification accuracy.

Dimensionality reduction methods learn a low dimensional embedding space which best preserves the original high dimensional feature space. For instance Golugula et. al. performed dimensionality reduction via supervised canonical correlation analysis to learn a low dimensional space in which patient classification was performed. Naik et. al. demonstrated that GE is well suited for the preservation of a high dimensional feature space which characterized histological differences in texture, nuclear architecture, and gland morphology. Similarly, DBS features in conjunction with GE resulted in 88% classification accuracy for Gleason pattern 3 and 4 glands. This suggests that there is a use for ML schemes in facilitating classification of high dimensional histopathologic data.

3.2 Manifold Regularization

ML is well known to be sensitive to the dataset considered, as well as noise and outliers contained within a dataset. Perturbations in the manifold structure may reduce classification performance in the low dimensional embedding space as object-class relationships may be obscured. Manifold regularization techniques have
been proposed which impose additional constraints on ML to better preserve object-class relationships in the low dimensional space. For instance, Chang et al. proposed a weighted ML scheme, where outlier samples were assigned low weights, to reduce the effect outliers have on learning the manifold.\(^27\) Other manifold regularizers perform local smoothing on the learned manifold.\(^29\) Manifold regularization techniques may add a smoothness constraint into the ML algorithm.\(^26,30\) All of these methods over smooth the manifold, as they reduce the effects of outliers which including meaningful information as well as noise.

Another type of regularization learns a consensus embedding (CE) from a set of manifolds. Hou et al. learned a set of manifolds by obtaining multiple views for each sample in the dataset and then generated a consensus manifold across the views.\(^31\) Other CE schemes have varied the parameters or samples considered to find a manifold set, and then generated a CE from the set.\(^32,33\) These methods rely on the manifolds in the set being independent, which may not be a valid assumption when generating manifold sets across ML parameters. Additionally, relationships between samples across the manifolds in the set are not taken into account when determining a CE.

In this work we present a statistical shape model of manifolds (SSMMs) to perform manifold regularization. The idea being that a Statistical Shape Model\(^28\) can be utilized to determine the statistical likelihood of a manifold shape utilizing a manifold set. The SSMM can be utilized in several ways. (1) Regions on a new, related manifold which deviate from the SSMM can be identified. By identifying these regions, meaningful differences between the dataset and the SSMM may be determined. (2) Noisy samples on a manifold can be identified based on their deviation from the SSMM. Removing these samples from the dataset may result in a more accurate low dimensional manifold, and hence improve classification accuracy. (3) A classifier can be trained on the SSMM which would allow for (a) classifier decision boundaries to be applied to a new, related manifold without retraining the classifier or (b) new, related samples could be projected onto the SSMM. The projection of newly acquired samples onto a previously calculated manifold can be performed by OSE.\(^34\)

4. BRIEF OVERVIEW AND NOVEL CONTRIBUTIONS

A flowchart of the proposed SSMM methodology is displayed in Figure 3. Table 1 list the notation used throughout the paper. To construct the SSMM we:

**Step 1: Construction of Uncorrelated Manifolds.** Generate a set of uncorrelated manifolds \(M\) for a dataset \(O\). For this task we divide the dataset \(O\) into \(K\) folds, and then generate \(M\) using a leave-one-fold-out scheme.

**Step 2: Procrustes based Registration.** As manifolds in \(M\) will be misaligned, primarily due to rotational and translational differences, a Procrustes based registration scheme is utilized to align the manifolds in \(M\).

**Step 3: Statistical Shape Manifold Model.** Calculate the mean manifold shape and primary modes of variation for \(M\).

**Step 4: Constraining a New Manifold Instance.** Once constructed the SSMM constrains a new manifold instance \(M\) of related datasets to only those shapes statistically most likely to occur resulting in the regularized manifold \(\tilde{M}\).

In this work we demonstrate that the SSMM can (a) determine noisy samples by identifying samples which deviate from the SSMM, and (b) accurately perform OSE of newly acquired samples onto a manifold constrained by the SSMM. The novel contributions of the SSMM are:

- The first ever application of SSMs to manifolds.
- A computerized decision support system which utilizes a SSMM based on the morphologic features of glands on prostate histopathology to automatically distinguish between Gleason patterns 3 and 4.
- A Procrustes based registration scheme to align all manifolds in \(M\). Each sample \(o_i \in O\) has a corresponding embedding location \(y_{i,k}\) on the manifold \(M_k\). Procrustes registration aligns all manifolds in \(M\) to minimize the differences between \(y_{i,k}\) for all \(k \in \{1,\ldots,K\}\) and all \(o_i \in O\).
5. REVIEW OF THE THEORY OF MANIFOLD LEARNING

In this section we provide a theoretical overview of the specific ML concepts we leverage in this paper to construct our uncorrelated set of manifolds. GE, a specific ML algorithm, is discussed in Section 5.1. The Nyström method (NM), a specific OSE algorithm, is discussed in Section 5.2.

5.1 Graph Embedding (GE)

In this work, we implemented the ML scheme GE\(^{23}\) to perform nonlinear dimensionality reduction as it has few parameters to optimize over (only \(\gamma\) an empirically determined scaling term) and is relatively computationally efficient. GE learns a manifold estimate \(\hat{M}\) for a dataset \(O_T \in \mathbb{R}^D\). \(\hat{M}\) is described by a set of embedding locations \(\hat{y} \in \mathbb{R}^n\) where \(n ≪ D\). \(\hat{y}\) is obtained by performing the eigenvalue decomposition (EVD),

\[
W\hat{y}' = \lambda D\hat{y}',
\]

where \(W(a, b) = e^{-A(a, b)/\gamma}\), \(\gamma\) is an empirically determined scaling term, and \(D\) is the diagonal matrix \(D(a, a) = \sum_b W(a, b)\). \(\hat{y}\) is defined as the \(n\) eigenvectors which correspond to the top \(n\) eigenvalues in \(\lambda\).

5.2 Out-of-Sample Extrapolation (OSE)

A sample not in the original dataset, i.e. \(o_k \notin O_T\), will not have a corresponding embedding location in \(\hat{y}\). To calculate the embedding location \(\hat{y}_k\) the dissimilarity matrix \(A\) and the EVD would have to be recomputed to include \(o_k\) in \(O_T\). Repeating ML for every new sample acquired is computationally infeasible.\(^{34}\) The aim of OSE is to determine embedding locations \(\hat{y}\) for samples in a newly acquired dataset defined as \(O_R\).
The Nyström Method (NM) is a OSE algorithm which estimates $\hat{y}$ as a weighted sum of the known embeddings $\hat{y}$. Given $\hat{y}$ for $O_T$ generated in Section 5.1, $\tilde{y}$ for $O_R$ are calculated as,

$$
\tilde{y}_{i,d} = \frac{1}{\hat{\lambda}_d} \sum_{j=1}^{S} \hat{y}_{j,d} W(o_j,o_i),
$$

where $d \in \{1, \ldots, n\}$ is the $d$th embedding dimension corresponding to the $d$th largest eigenvalue $\hat{\lambda}_d$.

**6. CONSTRUCTION OF STATISTICAL SHAPE MODEL OF MANIFOLDS (SSMM)**

**6.1 Construction of Uncorrelated Manifolds**

A set of $K$ manifolds $M = \{\hat{M}_1, \ldots, \hat{M}_K\}$ are obtained from a dataset of $N$ samples defined as $O = \{o_1, \ldots, o_N\}$. $M$ is generated utilizing a $K$ fold scheme via the following steps:

1. Samples in $O$ are randomly divided into $K$ equal partitions such that $O = \{O_1 \cup \ldots \cup O_K\}$.
2. Testing and training sets are obtained via a leave-one-fold-out scheme. A testing set is defined as $O_{R,k} = O_k : k \in \{1, \ldots, K\}$ and the corresponding training set is defined as $O_{T,k} \cup O_{R,k} = O$.
3. Each training set $O_{T,k}$ is utilized to find $\hat{M}_k$ which is defined as $\hat{y}_k$ calculated from GE as described in Section 5.1.
4. Each test set $O_{R,k}$ is extrapolated into the manifold $\hat{M}_k$ to determine $\tilde{y}_k$ via NM as described in Section 5.2.
5. Training and testing sets are combined to determine $y_k = \{\hat{y}_k, \tilde{y}_k\}$.

GE and NM were chosen for this application however any ML and OSE scheme can be used to construct $M$. 

---

**Table 1. Notation used in the paper.**

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$O$</td>
<td>Dataset</td>
<td>$</td>
<td></td>
</tr>
<tr>
<td>$o_i, o_j$</td>
<td>Samples contained in $O$</td>
<td>$M$</td>
<td>Aligned manifold set</td>
</tr>
<tr>
<td>$O_k$</td>
<td>$k$th fold of $O$ for $k \in {1, \ldots, K}$</td>
<td>$M$</td>
<td>Mean manifold shape for $\hat{M}$</td>
</tr>
<tr>
<td>$\mathbb{R}^D$</td>
<td>High dimensional feature space</td>
<td>$P$</td>
<td>Primary modes of variation for $\hat{M}$</td>
</tr>
<tr>
<td>$\psi(·,·)$</td>
<td>Dissimilarity measure</td>
<td>$\hat{M}$</td>
<td>New manifold instance</td>
</tr>
<tr>
<td>$A$</td>
<td>Dissimilarity matrix defined as $\psi(o_i,o_j)$ evaluated for all $o_i,o_j \in O$</td>
<td>$M$</td>
<td>Manifold constrained via the SSMM</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>GE scaling term</td>
<td>$\hat{Q}$</td>
<td>New dataset instance</td>
</tr>
<tr>
<td>$\mathbb{R}_n$</td>
<td>Low dimensional embedding space</td>
<td>$Q_n$</td>
<td>Samples which contain noise in $Q$</td>
</tr>
<tr>
<td>$M$</td>
<td>Uncorrelated manifold set</td>
<td>$Q_c$</td>
<td>Samples which do not contain noise in $Q$</td>
</tr>
<tr>
<td>$\hat{M}_k$</td>
<td>$k$th manifold in $M$</td>
<td>$\tau$</td>
<td>Threshold to determine sample deviation from the SSMM</td>
</tr>
<tr>
<td>$y_k$</td>
<td>Embedding locations on $\hat{M}_k$</td>
<td>$\hat{M}_c$</td>
<td>Manifold generated from $Q_c$</td>
</tr>
<tr>
<td>$y_{i,k}$</td>
<td>Embedding location for $o_i$ on $\hat{M}_k$</td>
<td>$Q_{te}$</td>
<td>Testing samples not contained in $Q$</td>
</tr>
<tr>
<td>$T^{a,b}$</td>
<td>Transformation to align $\hat{M}_b$ to $\hat{M}_a$</td>
<td>$\hat{M}_{te,c}$</td>
<td>Manifold with samples in $Q_{te}$ projected onto $\hat{M}_c$.</td>
</tr>
</tbody>
</table>
6.2 Manifold Alignment via Procrustes Based Registration

Manifolds contained in \( \mathbb{M} \) may not align, the algorithm for ML preserves pairwise relationships between samples but may not preserve the global relationship of samples in the low dimensional embedding space. Procrustes registration is applied to affinely align all manifolds in \( \mathbb{M} \). Procrustes registration can be performed since there are point correspondences between all manifolds in \( \mathbb{M} \) as each sample in \( \mathbb{O} \) has a location on every manifold in \( \mathbb{M} \).

A reference manifold \( \hat{\mathcal{M}}_a : a \in \{1, \ldots, K\} \) is randomly selected. All other manifolds, \( \hat{\mathcal{M}}_b : b \neq a \) are registered to \( \hat{\mathcal{M}}_a \) by minimizing,

\[
\hat{\mathcal{M}}_b = \min_{T_{a,b}} \left( \sum_i ||y_{i,a} - T_{a,b}(y_{i,b})|| \right),
\]

where a sample \( o_i \) has the embedding locations \( y_{i,a} \in \mathcal{M}_a \) and \( y_{i,b} \in \mathcal{M}_b \). \( ||\cdot|| \) denotes the L2-norm. Registration is performed for all \( \hat{\mathcal{M}} \in \mathbb{M} \) to obtain the aligned set of manifolds \( \hat{\mathcal{M}} \).

6.3 Statistical Shape Model of Manifolds

Once all manifolds are aligned the statistical properties of \( \hat{\mathcal{M}} \) can be determined. The SSMM is defined via the mean and principal modes of variation for \( \hat{\mathcal{M}} \). The mean of \( \hat{\mathcal{M}} \) is calculated by,

\[
\bar{\mathcal{M}} = \frac{1}{K} \sum_{k} \hat{y}_{i,k} : \forall \hat{y}_{i,k} \in \hat{\mathcal{M}}_k.
\]

The principal modes of variation for the manifold defined as \( P \) are obtained by performing PCA on \( \hat{\mathcal{M}} \). Only the \( P \) corresponding to the top 95% of variation are considered to constrain the SSMM to likely shapes.

6.4 Constraining a New Manifold Instance to the SSMM

A new manifold \( \hat{\mathcal{M}} \) is obtained from applying GE to \( \mathbb{O} \). \( \hat{\mathcal{M}} \) is constrained to only likely shapes as defined by the SSMM obtained in Section 6.3.

\[
\hat{\mathcal{M}} = T^{a,K+1} (\bar{\mathcal{M}} + P * b),
\]

where \( b \) controls the shape of \( \hat{\mathcal{M}} \) and \( T^{a,K+1} \) is an affine transformation between the SSMM and \( \hat{\mathcal{M}} \). \( b \) is constrained to \( \hat{\mathcal{M}} \pm 2\sigma \) to limit the SSMM to only those shapes statistically most likely to occur.

6.5 Application of SSMM to Identify Noisy Samples

The SSMM can aid in the identification of samples which contain noise. The algorithm \textit{FilterManifold} assumes samples which contain noise are those samples that deviate most from the SSMM. A dataset contains \( N \) samples defined as \( Q = \{q_1, \ldots, q_N\} \). The following algorithm can be used to identify the samples which contain noise \( Q_n \) and the samples which do not contain noise \( Q_c \) within \( Q \) given a user defined threshold \( \tau \).

**Algorithm FilterManifold**

**Input:** \( Q, \tau \)

**Output:** \( \mathcal{M}_c \)

**begin**

1. Obtain \( \hat{\mathcal{M}} \) from \( Q \) via application of the SSMM.
2. Obtain \( \mathcal{M} \) from \( Q \) by GE (Eq. 1).
3. Calculate \( e(q_i) = ||\hat{y}_i - \hat{\tilde{y}}_i|| \).
4. Obtain \( Q_n = q_i : q_i \in Q, e(q_i) \geq \tau \).
5. Obtain \( Q_c : Q_c \cup Q_n = \emptyset \).
6. Obtain \( \mathcal{M}_c \) for \( Q_c \) via GE (Eq. 1)

**end**
Dataset | Sample Size | Dissimilarity Measure
--- | --- | ---
Synthetic Swiss Roll\textsuperscript{22} | 3000 | $\psi(o_i, o_j) \begin{cases} ||o_i - o_j|| & \text{if } ||o_i - o_j|| < N, \\ 0 & \text{otherwise.} \end{cases}$ \(N\) is a neighborhood parameter.

Synthetic Helix\textsuperscript{22} | 3000 | $\psi(o_i, o_j) \begin{cases} ||o_i - o_j|| & \text{if } ||o_i - o_j|| < N, \\ 0 & \text{otherwise.} \end{cases}$ \(N\) is a neighborhood parameter.

Needle core prostate biopsy sections stained with H & E | 888 (58 patients) | Diffeomorphic Based Similarity (DBS)\textsuperscript{17}

Table 2. Description of datasets and their dissimilarity measures.

6.6 Application of SSMM to OSE

The SSMM can be utilized for robust OSE, by generating a more accurate manifold representation of a dataset. The algorithm OSE-SSMM demonstrates how the SSMM can be used for this purpose. A dataset \(Q\) is divided into training samples \(Q_{tr}\) and testing samples \(Q_{te}\) such that \(Q_{tr} \cup Q_{te} = \emptyset\). To find a set of testing embeddings \(M_{te,c}\) for a filtered manifold we apply the following algorithm,

Algorithm OSE-SSMM
Input: \(Q_{tr}, Q_{te}, \tau\)
Output: \(M_{te,c}\)
begin
1. Obtain \(Q_{tr,c}\) and \(Q_{tr,n}\) for \(Q_{tr}\) via FilterManifold.
2. Obtain \(M_{tr,c}\) for \(Q_{tr,c}\) via GE (Eq. 1).
3. Obtain \(M_{te,c}\) for \(Q_{te}\) via NM (Eq. 2) with \(M_{tr,c}\) as the training manifold.
end

7. EXPERIMENTAL DESIGN AND RESULTS

7.1 Dataset Description

7.1.1 Synthetic Datasets

Two synthetic datasets, Swiss Roll and Helix, described in Table 2 were utilized to demonstrate the application of SSMM to manifold regularization. The Swiss Roll is a 2D planar manifold divided into two classes which exists in a 3D space. The Helix is a 2D circular manifold divided into six classes which exists in a 3D space. The benefit of both datasets is that the high dimensional 3D space and the low dimensional 2D embedding space may be visualized. Gaussian noise was added to 2% of samples within each dataset where the standard deviation of the noise was set equal to 5% of the standard deviation of samples in the dataset. The dissimilarity measures for both datasets are reported in Table 2.

7.1.2 Prostate Histopathology

Prostate needle core tissue biopsies were obtained from 58 patients. Biopsies were stained with H & E and digitized at 40× optical magnification using an Aperio scanner. An expert pathologist selected regions of interest (ROIs) on each biopsy. In total 120 ROIs were selected across. Each ROI was assigned a Gleason pattern of either BE, G3, or G4. All glands contained within each ROI were manually segmented to obtain a total of 888 glands from BE (\(N = 93\)), G3 (\(N = 748\)), and G4 (\(N = 47\)) ROIs. DBS was the dissimilarity measure utilized to quantify morphologic differences between glands.\textsuperscript{17}
Table 3. (a) SI and (b) CA are reported for \( M \) and \( M_c \). The best value for each dataset is bolded. p-values are reported for a Student’s t-test comparing \( M \) and \( M_c \).

<table>
<thead>
<tr>
<th>Dataset</th>
<th>( M )</th>
<th>( M_c )</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Swiss Roll</td>
<td>0.56 ± 0.01</td>
<td>0.57 ± 0.03</td>
<td>0.063</td>
</tr>
<tr>
<td>Helix</td>
<td>0.44 ± 0.05</td>
<td>0.47 ± 0.02</td>
<td>0.138</td>
</tr>
<tr>
<td>Prostate</td>
<td>0.02 ± 0.01</td>
<td>0.05 ± 0.03</td>
<td>0.032</td>
</tr>
</tbody>
</table>

Table 4. (a) SI and (b) CA are reported for \( M_{te} \) and \( M_{te,c} \). The best value for each dataset is bolded. p-values are reported for a Student’s t-test comparing \( M_{te} \) and \( M_{te,c} \).

<table>
<thead>
<tr>
<th>Dataset</th>
<th>( M_{te} )</th>
<th>( M_{te,c} )</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Swiss Roll</td>
<td>0.57 ± 0.01</td>
<td>0.58 ± 0.01</td>
<td>0.061</td>
</tr>
<tr>
<td>Helix</td>
<td>0.47 ± 0.01</td>
<td>0.47 ± 0.01</td>
<td>0.77</td>
</tr>
<tr>
<td>Prostate</td>
<td>-0.04 ± 0.01</td>
<td>-0.02 ± 0.02</td>
<td>0.005</td>
</tr>
</tbody>
</table>

Table 7.2 Evaluation Measures

7.2.1 Silhouette Index (SI)

SI is a measure of how well samples cluster by class label\(^{36}\) with 1 corresponding to perfect clustering by class and -1 corresponding to no clustering by class. SI is calculated as, 

\[
\eta^M = \sum_{i=1}^{N} \frac{G(i) - C(i)}{\max[C(i), 0]}
\]

where 

\[
C(i) = \sum_{j, j \neq i} ||y_i - \bar{y}_j||
\]

and 

\[
G(i) = \sum_{j, j \neq i} ||\bar{y}_i - \bar{y}_j||
\]

7.2.2 Classification Accuracy (CA)

A probabilistic boosting tree (PBT) classifier\(^{37}\) was trained and evaluated using 3 fold cross validation. A dataset was divided into 3 folds such that all samples from a single patient were contained in the same fold and all folds maintained class balance.

7.3 Experiment 1: Application of SSMM to Filtered Manifold Learning

For each dataset \( Q \) in Table 2, a manifold \( M \) was calculated from \( Q \) using GE as described in Section 5.1. Similarly a filtered manifold \( M_c \) was found by FilterManifold as described in Section 6.5. The measures described in Section 7.2 were used to evaluate \( M \) and \( M_c \). A Student’s t-test was calculated to determine the statistical significance between \( M \) and \( M_c \) for each evaluation measure described in Section 7.2. Experimental results for all datasets are reported in Table 3. Across all datasets \( M_c \) outperforms \( M \) in terms of SI and CA. In the prostate histology dataset these increases in SI and CA were statistically significant. Hence \( M_c \) is better able to preserve object-class relationships in the datasets evaluated. For the synthetic datasets changes in SI and CA are not always statistically significant. However, as shown in Figure 1, qualitatively (d) \( M_c \) is a closer approximation to (b) the true embedding than (c) \( M \). In Figure 1 the samples are colored according to location on the true embedding to aid in visualization.

7.4 Experiment 2: Application of SSMM to Filtered OSE

For each dataset \( Q \) in Table 2, a training set \( Q_{tr} \) and a testing set \( Q_{te} \) were defined so that \( Q_{te} \) is 10% of \( Q \) and \( Q_{tr} \cup Q_{te} = \emptyset \). \( Q_{tr} \) and \( Q_{te} \) were used to construct an original manifold \( M_{te} \) and filtered manifold \( M_{te,c} \). \( M_{te} \) is generated by applying GE as described in Section 5.1 and then applying NM as described in Section 5.2 to \( Q_{te} \) where \( M_{te} \) is the training manifold. The filtered manifold \( M_{te,c} \) is calculated by OSE-SSMM as described in Section 6.6. The measures described in Section 7.2 were used to evaluate \( M_{te} \) and \( M_{te,c} \). A Student’s t-test was calculated to determine the statistical significance between \( M_{te} \) and \( M_{te,c} \) for each evaluation measure described in Section 7.2. Experimental results for all datasets are reported in Table 4. For the histopathology dataset \( M_{te,c} \) outperforms \( M_{te} \) in terms of SI and CA. The synthetic datasets, the Swiss Roll and Helix, do not show improved performance.
8. CONCLUDING REMARKS

In this paper we presented a statistical shape model of manifolds (SSMM) to perform manifold regularization. The SSMM models the low dimensional embedding space, found via ML, of a dataset with a statistical shape model (SSM). New, related manifolds may then be constrained by the SSMM to only those shapes statistically most likely to occur. The SSMM may be utilized for several applications including: (a) identification of samples which contain noise based on their deviation from the SSMM. Removing these samples from the dataset may result in a better low dimensional representation of the relationship between samples in the dataset. (b) A classifier could be trained on the SSMM allowing for (i) classifier decision boundaries to be applied to a new related manifold without retraining the classifier or (ii) new, related samples to be classified by projection of the samples onto the SSMM. (c) Identification of regions on a new, related manifold which deviate from the SSMM. Identifying these regions may aid in determining meaningful differences between the dataset and SSMM.

To construct the SSMM we (1) generate a set of uncorrelated manifolds \( M \) for a dataset \( O \), (2) align manifolds in \( M \), and (3) calculate the mean manifold shape and its primary modes of variation. The SSMM can constrain a new, related manifold instance to only those shapes statistically most likely to occur resulting in a regularized manifold. We have demonstrated that SSMM can improve classification accuracy (CA) in the context of Gleason grading of prostate histopathology utilizing quantitative morphologic features of glands. For the dataset considered, histology belonged to either benign, Gleason pattern 3, or pattern 4 tissue classes. Improvements in CA via the SSMM were demonstrated for two applications: (a) We demonstrated that outlier samples within a manifold can be identified as those samples which deviate from the SSMM. Removal of outlier samples increased CA and SI. (b) We demonstrated that manifold regularization by the SSMM improves SI CA when performing out-of-sample extrapolation (OSE) of never before seen samples onto the SSMM.

In future work we intend to explore the ability of the SSMM to uncover meaningful differences between a new, related dataset and the SSMM. We plan to research the ability of the SSMM to identify regions of a new, related manifold which deviate from the SSMM. These regions will then be further investigated to determine subtle difference between the dataset and the SSMM.

9. ACKNOWLEDGEMENTS

This work was made possible via grants from the Wallace H. Coulter Foundation, National Cancer Institute (Grant Nos. R01CA136535-01, R01CA14077201, and R03CA143991-01), the Cancer Institute of New Jersey, the Department of Defense (W81XWH-11-1-0179), and the Society for Imaging Informatics in Medicine (SIIM). We would like to thank Dr. J. E. Tomaszewski from the University at Buffalo School of Medicine and Biomedical Sciences as well as Drs. M. D. Feldman and Natalie Shi from the Hospital of the University of Pennsylvania for providing prostate histology imagery and corresponding annotations.

REFERENCES


