CONTENT-BASED IMAGE RETRIEVAL OF DIGITIZED HISTOPATHOLOGY VIA BOOSTED SPECTRAL EMBEDDING (BoSE)

BY AKSHAY SRIDHAR

A thesis submitted to the

Graduate School—New Brunswick

Rutgers, The State University of New Jersey

in conjunction with

The Graduate School of Biomedical Sciences

the University of Medicine and Dentistry of New Jersey

in partial fulfillment of the requirements for the

Joint Degree of Master of Science

Graduate Program in Biomedical Engineering

Written under the direction of

Dr. Anant Madabhushi

and approved by

_____

_____

_____

_____

New Brunswick, New Jersey

January, 2012

ABSTRACT OF THE THESIS

Content-Based Image Retrieval of Digitized Histopathology via Boosted Spectral Embedding (BoSE)

by Akshay Sridhar

Thesis Director: Dr. Anant Madabhushi

Content-based image retrieval (CBIR) systems allow for retrieval of images from a database that are similar in visual content to a query image. This is particularly useful in scenarios such as digital pathology, where text-based descriptors alone might be inadequate to accurately describe image content. By representing images via a set of quantitative image descriptors, the similarity between a query image with respect to archived, annotated images in a database can be computed and the most similar images retrieved. Recently, non-linear dimensionality reduction (NLDR) methods have become popular for embedding high dimensional data into a reduced dimensional space while preserving local object adjacencies, thereby allowing for object similarity to be determined more accurately in the reduced dimensional space. However, most dimensionality reduction (DR) methods implicitly assume, in computing the reduced dimensional representation, that all features are equally important. Erroneous or noisy features could potentially result in dissimilar images being mapped close to each other in the reduced embedding space. In this work we present Boosted Spectral Embedding (BoSE), a variant of the traditional Spectral Embedding (SE) NLDR method, which unlike SE utilizes a boosted distance metric (BDM) to selectively weight individual features to subsequently map the data into a reduced dimensional space. In this work BoSE is evaluated against SE (which employs equal feature weighting) in the

context of CBIR of digitized prostate and breast cancer histopathology images. Across 154 hematoxylin and eosin (H&E) stained histopathology images corresponding to benign and malignant prostate cancer biopsy images, low and high grade ER+ breast cancer studies, and HER2+ breast cancer H&E images, BoSE outperformed SE both in terms of CBIR-based (area under the precision recall curve) and classifier-based (classification accuracy) performance measures. Consistent trends were observed when embedding the data into spaces with different dimensions. Our results suggest that BoSE could serve as an important tool for CBIR and classification of high dimensional biomedical data.

Acknowledgements

I would like to first thank my advisor Dr. Anant Madabhushi. He took a chance on me when no other lab would. He has since then both challenged and inspired me to achieve what I never thought was possible. His dedication to the members of the lab has shown me what true leadership is all about.

I would also like to thank the members of the Laboratory for Computations Imaging and Bioinformatics. They created a wonderful working environment with their willingness to help each other with anything despite their work load. I would specifically like to thank Dr. Scott Doyle and Ajay Basavanhally for their contributions to this work.

Lastly, I would like to thank my parents without whom none of this would be possible. They have always pushed me to achieve excellence and provided me with the tools to do so.

This work was made possible by the Wallace H. Coulter Foundation, New Jersey Commission on Cancer Research, National Cancer Institute (R01CA136535-01, R01CA140772-01, and R03CA143991-01), The Cancer Institute of New Jersey, and The Hospital at the University of Pennsylvania. I also wish to thank Dr. John Tomaszewski, Dr. Michael Feldman, Dr. Natalie Shih, Dr. Carolyn Mies, and Dr. Shridar Ganesan for providing and annotating the digitized histopathology data.

Dedication

"Mom: Someday I hope you have a kid that puts you through what I've gone through. Calvin: Yeah, Grandma said that's what she used to tell you."
– Bill Watterson

This work is dedicated to my parents. Thank you for not giving up on me despite all of my ridiculousness.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background and Motivation

Content-based image retrieval (CBIR) systems allow a user to retrieve images from a database based on visual similarity to the query image. This is particularly useful for digital pathology and medical imaging databases, where text-based descriptors alone might be inadequate to accurately describe image content [2, 3, 4, 5, 6, 7, 8]. In CBIR systems, a query image is used as the input and based on image attribute matching, the most similar images from within a database are retrieved. All images are represented by a unique set of numbers termed features that describe various aspects of the images. Two main components of a CBIR system are (a) the image (or feature) representation, and (b) choice of similarity metric for performing retrieval. An ideal similarity metric would yield a large value when comparing visually dissimilar images and a small value when similar images are compared. For any given query image, the most similar images in the database as determined by the similarity metric are retrieved in decreasing order of relevance. However, in cases where images are represented by a large number of image attributes, the similarity measure might be affected by the so called "curse of dimensionality" problem, wherein the number of attributes may be greater than the total number of instances in the database.

Dimensionality reduction (DR) is a technique that is used to project high dimensional data into a reduced dimensional embedding space. The low dimensional data representation allows for more consistent and accurate similarity computations, compared to the high dimensional space, to help determine image similarity [9][10]. DR techniques can be broadly categorized as linear or nonlinear. Linear DR techniques such as principal component analysis (PCA) [11] fail to accurately capture object (image) relationships where the data resides on some non-linear manifold [12]. Objects residing on different ends of the manifold could

potentially be mapped closer to each other in the lower dimensional space, since linear DR methods use the Euclidean norm as opposed to the geodesic distance (appropriate for adjacency determination for objects residing on nonlinear manifolds). Nonlinear dimensionality reduction (NLDR) methods [13, 14, 15, 16] attempt to capture object adjacency on nonlinear manifolds by preservation of the local linear neighborhood structure [17]. However, NLDR methods such as Isomaps [13] and Locally Linear Embedding (LLE) [14] are sensitive to the choice of the size of the local neighborhood (κ) within which linearity is assumed. Diffusion Maps [15], another NLDR method, is sensitive to the number of time steps specified for the random walk. Spectral Embedding (SE) [16] is a NLDR method that, unlike neighborhood preserving NLDR schemes (such as LLE, Isomaps), defines object adjacency by using a Gaussian kernel in conjunction with the Euclidean distance metric (EDM) to yield a similarity matrix for all objects. The eigenvalue decomposition of this similarity matrix is then determined to yield the low dimensional representation (eigenvectors) of the data. While SE is still sensitive to the parameters of the kernel, it has been shown to be more robust compared to LLE and Isomaps [18]. CBIR could be performed in conjunction with SE by mapping the query and database images into a reduced dimensional space and then retrieving relevant images as those in the neighborhood of the query instance. A key shortcoming of the EDM, however, is that it implicitly assumes all features (dimensions) are equally relevant. In the context of CBIR, features that are poor in discriminating between two image classes could potentially map dissimilar images close to each other in the low dimensional space. Hence, in order to determine the optimal low dimensional representation of the data, it is desirable to weight the discriminatory attributes higher compared to the erroneous or noisy features prior to computing the similarity matrix.

1.2 Previous Work

1.2.1 Spectral Embedding Variants

There has been some previous work in the development of SE variants. Tiwari, et al. proposed a weighted multi-kernel learning scheme to yield an improved weight matrix for use in conjunction with SE [19]. ElGhawalby, et al. [20] formulated a variant of SE that used

an edge-based wave kernel that embedded the nodes of a graph as points on the surface of a manifold, and used the resulting point-set to compute graph characteristics. Robles-Kelly, et al. [21] used the Kruskal coordinates to compute the edge-weights for a weight matrix and used it to embed the nodes of the graph onto a Riemannian manifold.

1.2.2 Nonlinear Dimensionality Reduction with Content-Based Image Retrieval

NLDR schemes have previously been coupled with CBIR. Doyle, et al. [2] found that the choice of feature space greatly affected the ability of the CBIR system to output images of the same class as the query image. He, et al. [9] developed a dimensionality reduction scheme called Maximum Margin Subspace (MMP) that maximizes the margin between positive and negative samples at each local neighborhood. They projected the images into a lower dimensional space and retrieval was performed. Huang, et al. [10] proposed a method of representing images by treating them as frequency histograms of salient features and performed image retrieval in a lower dimensional space created by LDA.

CBIR has also been applied to various domains including medical images [3] [4]. In particular, it has been used to retrieve lung images [5], dermatological images [6], and histopathology [7] [8]. However, the retrieval of medical images has not been done in a learned reduced dimensional space.

1.3 Brief Overview of CBIR System

In this work we employ a novel variant of SE called Boosted Spectral Embedding (BoSE), a supervised NLDR technique that utilizes a boosted distance metric (BDM) in place of the EDM. The BDM, which was first introduced in [22], employs AdaBoost [23]. The AdaBoost [23] algorithm, a classifier ensemble, introduced by Freund and Schapire, allows for implicit feature weighting based on class discriminability. The difference between SE and BoSE is that BDM actively places importance on discriminatory features while mitigating the role of weaker features, yielding an embedding which encourages same class objects to be embedded closer to each other and dissimilar class objects to be mapped farther apart.

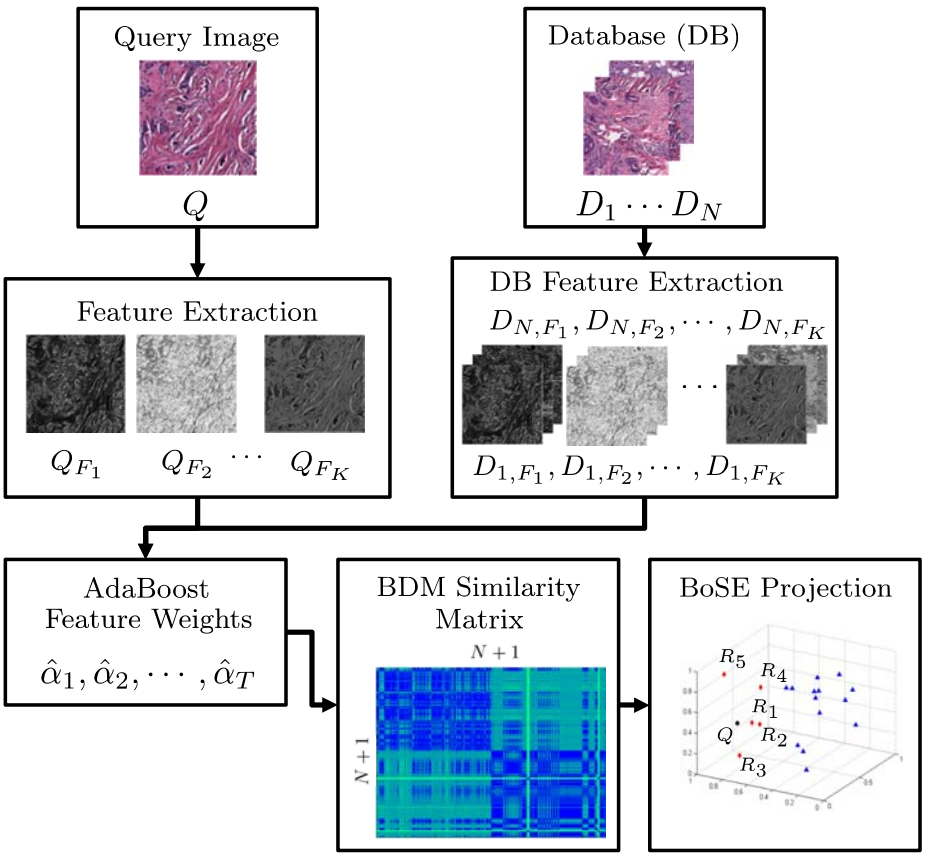


Figure 1.1: A flowchart illustrating the different components of the CBIR-BoSE system. Initially a query image Q is inputted and is followed by quantitative feature extraction to yield a set of K image descriptors $F_1, \ldots, F_K$. The database contains N annotated images (with corresponding class labels) with their corresponding feature-based representations. For the particular problem of interest, the image features are assigned weights $(\hat{\alpha}_1, \ldots, \hat{\alpha}_T)$ corresponding to their class separability. A weighted similarity matrix is then created via the BDM, which is then used with BoSE to project the data into a lower dimensional space. In the reduced space, the distance between the query Q and the database images is calculated and the database images most similar to the query are retrieved $(R_1, \ldots, R_5)$.

The primary contributions of this work are twofold. First we present a new NLDR scheme (BoSE) that employs AdaBoost with SE to generate lower dimensional data representations with greater class separability. Second, BoSE is employed in conjunction with a CBIR scheme (CBIR-BoSE) to perform accurate retrieval of database images with respect to a query instance. An overview of the CBIR-BoSE system is illustrated in Figure 1.1. For a database of N annotated images, feature extraction is performed to yield N corresponding high-dimensional feature vectors. A low dimensional embedding of this data ($\mathcal{M}^{\text{BoSE}}$) is then created via BoSE. The Euclidean distance between the query image and the database

images is then computed in $\mathcal{M}^{\text{BoSE}}$ and the most similar (lowest distance) images are first retrieved. Images retrieved from the same class as the query instance are considered as "relevant". Evaluation is done by constructing precision-recall (PR) curves, where a large area under the PR curve (AUPRC) reflects that CBIR-BoSE is retrieving the most relevant images first. A list of notation used in the thesis can be found in Table 1.1.

In this work we evaluated our CBIR-BoSE system on three different two class problems, illustrated in Figure 1.2. The three datasets comprised (1) 58 hematoxylin and eosin (H&E) stained prostate cancer tissue biopsy samples classified as benign (Figure 1.2 (a)) or malignant (Figure 1.2 (d)); (2) 55 H&E stained ER+ breast cancer histology specimens classified as low (Figure 1.2 (b)) or high (Figure 1.2 (e)) grade; and (3) 41 H&E stained HER2+ breast cancer tissue specimens classified as having low (Figure 1.2 (c)) or high (Figure 1.2 (f)) levels of lymphocytic infiltration (LI). The choice of these datasets was dictated by the fact that manual inspection of both prostate and breast cancer histology suffers from high inter- and intra-pathologist variability [24, 25, 26]. Typically the pathologist first determines if the histology sample is benign or malignant. If it is found to be malignant, the cancer is assigned a grade based on the morphologic and architectural attributes; cancer grade being highly correlated to patient outcome [24][27]. In the progression of solid tumors, local and systemic inflammation tends to play an important role [28]. Tumor infiltrating lymphocytes represent a local immune response and the degree of LI in a tumor is considered as being prognostic of patient outcome in several different disease states [29, 30, 31].

Symbol	Description
$\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$	Quantitative representation of images in $\mathbb{R}^{N \times D}$
$\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N\}$	Low dimensional projection of $\mathbf{X}$
$\mathbf{W}$	Weight matrix
Φ_d	Feature operator that extracts quantitative feature d from image
$\mathcal{L}(\mathbf{x}_i) \in \{+1, -1\}$	Ground truth label for object $\mathbf{x}_i$
h_d	Weak classifier built using a Bayesian framework
α_t	Weights associated with the t most optimal features
$\hat{\alpha}_t$	Normalized weights associated with the t most optimal features
$\mathbb{D}_{\text{BDM}}$	Boosted distance metric
$\mathcal{M}^{\text{BoSE}}$	Low dimensional representation produced by BoSE
$\mathcal{M}^{\text{SE}}$	Low dimensional representation produced by SE

Table 1.1: List of mathematical symbols and notations used throughout the thesis.

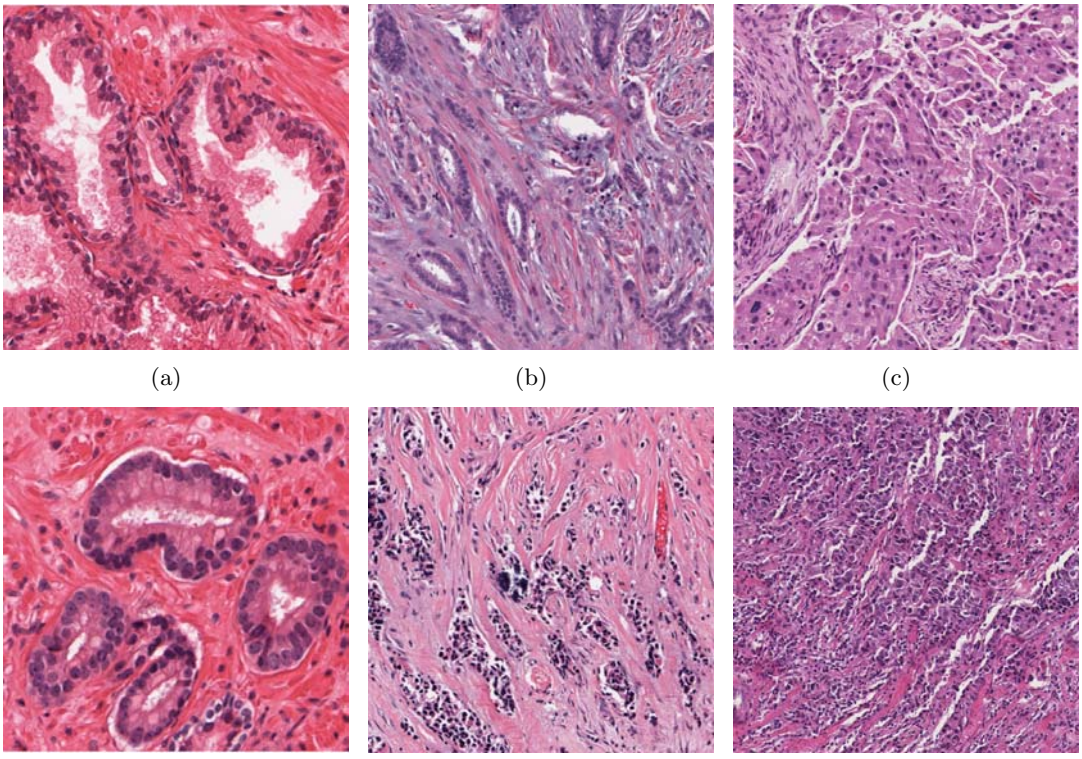


Figure 1.2: Example images of (a) benign and (d) malignant prostate tissue, (b) low and (e) high grade ER+ breast cancer tissue, and HER2+ breast cancer tissue with (c) low and (f) high levels of lymphocytic infiltration. The histology images were obtained by digitizing biopsy samples previously stained with hematoxylin and eosin (H&E). In (a) the nucleoli are less prominent and the glands are more open, whereas in (d) the nucleoli are more apparent and the glands are shriveled due to increased cell proliferation. There is a greater amount of nuclear proliferation in (e) high grade ER+ breast cancer when compared to (b). A similar phenomenon can be observed when looking at HER2+ breast cancer tissue with low vs. high levels of LI. In (f) there are more lymphocytes that have infiltrated the cancerous tissue compared to (c).

The development of CBIR tools with applications in digital pathology [32] could assist pathologists by providing a quantitative, reproducible and accurate image based risk score, indicative of disease aggressiveness and patient outcome [24]. Additionally, a CBIR system for digitized histopathology could serve as a teaching, training, and instructional tool for pathology residents and fellows.

1.4 Organization of Thesis

The rest of the thesis is organized as follows. The BDM is presented in Chapter 2. The methodological description of the BoSE scheme is presented in Chapter 3. The experimental design and evaluation of BoSE are presented in Chapter 4. Results and discussion are presented in Chapter 5. Lastly, concluding remarks are presented in Chapter 6.

Chapter 2

The Boosted Distance Metric

2.1 A Brief Overview of the Boosted Distance Metric (BDM)

We define a set of objects as $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ where N is the number of objects. Each image $\mathbf{x}_i, i \in \{1, \ldots, N\}$ belongs to one of two classes $+1$ or -1. The ground truth label of $\mathbf{x}_i$ is denoted $\mathcal{L}(\mathbf{x}_i) \in \{+1, -1\}$ where $\mathcal{L}(\mathbf{x}_i) = -1$ indicates membership in class -1 and $\mathcal{L}(\mathbf{x}_i) = 1$ indicates membership in class $+1$. Let $\Phi_d(\mathbf{x}_i)$ for $d \in \{1, 2, \ldots, D\}$ represent the value of feature d from $\mathbf{x}_i$. The BDM construction is comprised of three main steps:

Step 1: Constructing Weak Classifiers: Weak classifier $h_d(\mathbf{x}_i) \in \{-1, 1\}$ predicts the class label of $\mathbf{x}_i$ based on feature operator Φ_d. In this work, a weak classifier is one that outputs a class label for the object under consideration. The weak learner may be one that outputs a probabalistic likelihood that an object (in this case, an image) belongs to a specific class based solely on a single attribute. These probabilities can be thresholded to obtain the class label. Multiple different weak learners derived from various image features can be constructed and evaluated in terms of classifier accuracy (assuming that a training set with class labels is available). Weak classifiers were constructed by using only a subset (training set) of the entire dataset.

Step 2: Implicit Feature Weighting: The T most accurate weak classifiers, $h_t, t \in \{1, 2, \ldots, T\}$ are identified and weights $\hat{\alpha}_t$ associated with each h_t are learned via the AdaBoost [23] algorithm, thereby enabling implicit feature weighting.

Step 3: BDM Construction: The BDM is then defined using the features $\Phi_t(\mathbf{x}_i)$ and associated weights $\hat{\alpha}_t$ obtained in Step 2.

2.2 The Construction of Weak Classifiers

Each individual feature (weak classifier) is used to classify an image and its classification accuracy is leveraged in determining its class separability. The construction of the weak classifiers employed in this work is outlined below:

Step 1: Calculate $\Phi_d(\mathbf{x}_i)$ for all $d \in \{1, 2, \ldots, D\}, i \in \{1, 2, \ldots, N\}$, in order to obtain corresponding feature values for each of the images.

Step 2: Create training set $\mathbf{X}^{\text{tr}} \subset \mathbf{X}$ containing N objects by randomly sampling half of the entire dataset $\mathbf{X}$.

Step 3: Let $\mathbf{X}_+$ indicate all objects in $\mathbf{X}^{tr}$ belonging to class $+1$. Similarly, $\mathbf{X}_-$ is the set of all samples in $\mathbf{X}^{tr}$ that belong to class -1. We can obtain an appropriate probability distribution function (PDF) which predicts the likelihood of observing a feature value given a class label as:

$$p(\Phi_d(\mathbf{X}_a)|\omega_b) = \Phi_d(\mathbf{X}_a)^{\tau-1} \frac{\exp(\frac{-\Phi_d(\mathbf{X}_a)}{\eta})}{\eta^\tau \Gamma(\tau)}, \tag{2.1}$$

for $a \in \{+, -\}$, $\omega_b \in \{+1, -1\}$, Γ is the gamma function, and $\tau, \eta > 0$ are scale and shape parameters. Equation 2.1 is a gamma function estimation of the PDF [33], and is preferred to a Gaussian distribution because the feature histograms are asymmetric about the mean and the gamma function models the distribution more accurately.

Step 4: Obtain the *a posteriori* probability $P(+1|\Phi_d(\mathbf{x}_i))$ which computes the likelihood that an object with feature value $\Phi_d(\mathbf{x}_i)$ belongs to the positive class $+1$ by solving,

$$P(+1|\Phi_d(\mathbf{x}_i)) = \frac{P(+1)p(\Phi_d(\mathbf{x}_i)| + 1)}{P(+1)p(\Phi_d(\mathbf{x}_i)| + 1) + P(-1)p(\Phi_d(\mathbf{x}_i)| - 1)}. \tag{2.2}$$

Step 5: Once the *a posteriori* probabilities have been computed for each image based on a single feature, the weak classifiers are defined based off the individual features. The weak classifiers may now be defined as follows

$$h_d(\mathbf{x}_i) = \begin{cases} 1 & \text{if } P(+1|\Phi_d(\mathbf{x}_i)) > P(-1|\Phi_d(\mathbf{x}_i)) \\ -1 & \text{otherwise} \end{cases}$$

Algorithm: *BoostFeatWeights*
Input: Training samples $\mathbf{X}^{\text{tr}}$, ground truth labels $\mathcal{L}(\mathbf{X}^{\text{tr}})$, iterations T, weak classifiers h_d for $d \in \{1, 2, \ldots, D\}$
Output: Optimal classifiers h_t and their corresponding weights $\hat{\alpha}_t$
begin
0. Initialize distribution for samples $\Pi_1(i) = \frac{1}{N}$
1. **for** $t = 1$ to T
2. Find $h_t = \arg\min_{h_d}[\epsilon_d]$, where $\epsilon_d = \sum_{i=1}^{N} \Pi_t(i)[\mathcal{L}(\mathbf{x}_i) \neq h_d(\mathbf{x}_i)]$ for $\mathbf{x}_i \in \mathbf{X}^{\text{tr}}$;
3. *if* $\epsilon_t \geq 0.5$ *then* stop;
4. $\alpha_t = \frac{1}{2} \ln \frac{1-\epsilon_t}{\epsilon_t}$;
5. Update, $\Pi_{t+1}(i) = \frac{1}{Z_t} \Pi_t(i) \exp(-\alpha_t \mathcal{L}(\mathbf{x}_i) h_t(\mathbf{x}_i))$ for all $\mathbf{x}_i \in \mathbf{X}^{\text{tr}}$, where $Z_t = \sum_i \Pi_t(i) \exp(-\alpha_t \mathcal{L}(\mathbf{x}_i) h_t(\mathbf{x}_i))$ is a normalization term;
6. **endfor**
7. Normalize α_t to obtain $\hat{\alpha}_t$ such that $0 < \hat{\alpha}_t \leq 1$, $\hat{\alpha}_t = \frac{\alpha_t}{\max_t[\alpha_t]}$ for $t \in \{1, \ldots, T\}$.
8. **return** $\hat{\alpha}_t$ and h_t;
end

Figure 2.1: The *BoostFeatWeights* algorithm for implicitly weighting the top performing image features for a specific task. All samples were initialized with equal weights. The weights for the weak classifiers are computed based on the classification error ϵ_d. At each iteration, weights ($\Pi_t(i)$) increase for samples that are difficult to classify. This forces the weak classifiers to concentrate on the images that are frequently misclassified. Once all the weights (α_t) for the weak classifiers are found, the weights are normalized so that they would range from 0 to 1. The T best performing classifiers and their weights are outputted.

If the probability, which based on a single feature, of the image $\mathbf{x}_i$ belonging to class $+1$ is greater than its probability of belonging to class -1, it will be given a class label of 1. Otherwise, it will be given a classification label of -1.

2.3 Implicit Feature Weighting

We use the AdaBoost [23] algorithm to perform implicit weighting of the weak classifiers (in turn reflecting the importance of the individual image attributes) in order to distinguish between the positive and negative classes. Our feature weighting algorithm is illustrated in Figure 2.1. AdaBoost works in an iterative fashion by first identifying the best-performing weak classifiers and then assigning weights based on the discriminability of that feature [23]. The weights of the training images are initialized by taking the reciprocal of the number of images there are in the training set (Line 0). For each weak classifier (feature), its classification error is computed (Line 2). At each iteration, the weak classifier with the

lowest classification error is chosen and its weight is determined (Line 4). The weights of the training images are updated such that the images that were frequently classified properly recieved lower weights, while the images that were frequently misclassified recieved higher weights (Line 5). This ensures that subsequent weak classifiers are picked based on their ability to classify these hard to classify instances. The process repeats for T iterations. The output of the algorithm is a set of weak classifiers h_t and their associated normalized weights $\hat{\alpha}_t, t \in \{1, 2, \ldots, T\}$ where $1 \leq T \leq D$ and $0 < \hat{\alpha}_t \leq 1$. $\hat{\Phi}_t$ is the operator for the feature selected at iteration t of AdaBoost. The algorithm stops when $\epsilon_t > 0.5$.

2.4 Constructing the BDM

The BDM is constructed after the weights and features have been chosen. To find the distance between two points in the high dimensional space, we calculate,

$$\mathbb{D}_{\text{BDM}}(\mathbf{x}_i, \mathbf{x}_j) = \left[\sum_{t=1}^{T} \hat{\alpha}_t (\Phi_t(\mathbf{x}_i) - \Phi_t(\mathbf{x}_j))^2 \right]^{\frac{1}{2}}. \tag{2.3}$$

This is essentially a weighted Euclidean distance, where the weights influence the contribution of each feature. If $\hat{\alpha}_t \approx 0$, then Φ_t will not affect the value of the similarity measure.

2.4.1 Propositions for the BDM

Proposition 2.4.1 *Given that* $\mathbb{D}_{Eu} = \left[\sum_{t=1}^{T} (\Phi_t(\boldsymbol{x}_i) - \Phi_t(\boldsymbol{x}_j))^2 \right]^{\frac{1}{2}}$ *is the Euclidean distance metric,* $\mathbb{D}_{BDM}$ *is also a distance metric.*

Proof Since $\mathbb{D}_{\text{Eu}}$ is a metric, it is (1) positive, (2) symmetric, (3) definite, and (4) the triangle inequality holds. $\mathbb{D}_{\text{BDM}}$ must also be a metric since $\hat{\alpha}_t \in \mathbb{R}^+$ is positive and real valued. Therefore properties (1)-(4) are satisfied for $\mathbb{D}_{\text{BDM}}$. ∎

Proposition 2.4.2 below provides some insight into $\mathbb{D}_{\text{BDM}}$ for the simple case where $T = 2$, and where $a, b \in \mathbb{R}^2$.

Proposition 2.4.2 *If* $\mathcal{L}(a) = \mathcal{L}(b)$ *then* $\mathbb{D}_{Eu}(a, b) > \mathbb{D}_{BDM}(a, b)$.

Proof Ideally, if $\mathcal{L}(a) = \mathcal{L}(b)$, then $\mathbb{D}(a,b) \approx 0$. We denote the distance between a and b in the first dimension as Δ_1 and the second dimension as Δ_2. Assume that feature dimension Δ_1 is more discriminating than Δ_2; more specifically that $||\delta_1(a) - \delta_1(b)|| < ||\delta_2(a) - \delta_2(b)||$ where δ_1 and δ_2 represent the positions of the objects in feature spaces Δ_1 and Δ_2, respectively. Thus, $\hat{\alpha}_1 > \hat{\alpha}_2$ via the learned feature weights. Recall that $\mathbb{D}_{\text{BDM}}(a,b) = \sqrt{\hat{\alpha}_1(\Delta_1)^2 + \hat{\alpha}_2(\Delta_2)^2}$ and $\mathbb{D}_{\text{Eu}}(a,b) = \sqrt{(\Delta_1)^2 + (\Delta_2)^2}$. If the proposition is true, the following holds:

$$\sqrt{(\Delta_1)^2 + (\Delta_2)^2} > \sqrt{\hat{\alpha}_1(\Delta_1)^2 + \hat{\alpha}_2(\Delta_2)^2} \tag{2.4}$$

$$(\Delta_1)^2 + (\Delta_2)^2 > \hat{\alpha}_1(\Delta_1)^2 + \hat{\alpha}_2(\Delta_2)^2 \tag{2.5}$$

$$(\Delta_1)^2 - \hat{\alpha}_1(\Delta_1)^2 > \hat{\alpha}_2(\Delta_2)^2 - (\Delta_2)^2 \tag{2.6}$$

$$(\Delta_1)^2(1 - \hat{\alpha}_1) > (\Delta_2)^2(\hat{\alpha}_2 - 1) \tag{2.7}$$

Recall that $\hat{\alpha}_1, \hat{\alpha}_2 \geq 0$ and $\hat{\alpha}_1, \hat{\alpha}_2 \in [0,1]$. Therefore, the left hand side of the inequality would yield a positive number and the right hand side would yield a negative number, indicating that Proposition 2.4.2 holds. Note that it is similarly possible to show that under the same assumptions made for Proposition 2.4.1 if $\mathcal{L}(a) \neq \mathcal{L}(b)$, then $\mathbb{D}_{\text{Eu}}(a,b) < \mathbb{D}_{\text{BDM}}(a,b)$. ∎

Chapter 3

Boosted Spectral Embedding (BoSE) for Content-Based Image Retrieval

3.1 Boosted Spectral Embedding

The goal of SE is to project the feature vectors from a D dimensional space to a k dimensional space, where $k << D$. The low-dimensional representation of $\mathbf{X}$ is denoted $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N\}$. The first step in SE is to create a weight matrix $\mathbf{W}$, where each element (i, j) in $\mathbf{W}$ is denoted by w_{ij} and represents the distance between $\mathbf{x}_i$ and $\mathbf{x}_j$ defined by some metric $\mathbb{D}$.

The low dimensional representation of $\mathbf{X}$ is then found by solving the eigenvalue decomposition problem

$$(L - \mathbf{W})\mathbf{Y} = \lambda L\mathbf{Y}, \tag{3.1}$$

where L is the diagonal matrix, $L_{ii} = \sum_i w_{ij}$ [16].

The typical formulation of $\mathbf{W}$ involves the use of the EDM, where $w_{ij} = \exp(-\mathbb{D}_{\text{Eu}}(\mathbf{x}_i, \mathbf{x}_j)/\sigma)$, and σ is the standard deviation of $\mathbf{X}$. However, in BoSE, we replace the EDM with the BDM to obtain,

$$w_{ij} = \exp\Big(-\frac{\mathbb{D}_{\text{BDM}}(\mathbf{x}_i, \mathbf{x}_j)}{\sigma}\Big). \tag{3.2}$$

Since SE seeks to preserve object adjacencies as defined by $\mathbf{W}$, by improving the description of adjacency via the BDM, we should improve the resulting low dimensional embedding (achieve greater class separability in the reduced embedding space). Since $\mathbb{D}_{\text{BDM}}$ is a metric, $\mathbf{W}$ is positive, semi-definite, and symmetric.

Algorithm: *BoSE*
Input: Training samples $\mathbf{X}^{\text{tr}}$, Testing samples $\mathbf{X}^{\text{te}}$, $\mathcal{L}(X^{\text{tr}})$, $\mathcal{L}(X^{\text{te}})$, iterations T
Output: Lower dimensional embedding $\mathbf{Y}$
begin
1. Build weak classifiers $h_d : d \in \{1, 2, \ldots, D\}$ via a Bayesian Classifier;
2. Select optimal weak classifiers h_t and weights $\hat{\alpha_t}$ for $t \in \{1, 2, \ldots, T\}$ via AdaBoost;
3. Obtain BDM by applying Equation 2.3;
4. Obtain $\mathbf{W}$ by Equation 3.2;
5. Find $\mathbf{Y} \in \mathbb{R}^{N \times k}$;
6. **return** $\mathbf{Y}$

end

Figure 3.1: The BoSE algorithm. The weak classifiers are built using the training samples ($\mathbf{X}^{\text{tr}}$) and the weights are calculated via AdaBoost. The BDM is then employed with the weights to calculate the distances between all the objects in $\mathbf{X}$. The distances are used in conjunction with the Gaussian kernel to obtain the weight matrix $\mathbf{W}$. The lower dimensional embedding $\mathbf{Y}$ is then obtained by solving the eigenvalue decomposition in Equation 3.1.

Algorithm: *CBIR-BoSE*
Input: Query image $\mathbf{Q}$, database images $\mathbf{X}^{db} \in \mathbb{R}^{N \times D}$
Output: Top $\mathcal{N}$ Retrieved Images
begin
1. Calculate $\mathbf{x}^{query} = \Phi_d(\mathbf{Q})$ for all $d \in \{1, 2, \ldots, D\}$, where $\mathbf{x}^{query} \in \mathbb{R}^{1 \times D}$;
2. Concatenate $\mathbf{x}^{query}$ with $\mathbf{X}^{db}$ to form $\mathbf{X}^{all} \in \mathbb{R}^{(N+1) \times D}$;
3. Input $\mathbf{X}^{all}$ into BoSE to yield $\mathbf{Y}^{all} \in \mathbb{R}^{(N+1) \times k}$ where $k << D$;
4. Extract reduced query vector from $\mathbf{Y}^{all}$ to yield $\mathbf{y}^{query} \in \mathbb{R}^{1 \times k}$ and $\mathbf{Y}^{db} \in \mathbb{R}^{N \times k}$;
5. Calculate $\mathbf{p} = \mathbb{D}_{\text{Eu}}(\mathbf{y}^{query}, \mathbf{Y}_i^{db}), i \in \{1, 2, \ldots, N\}, \mathbf{p} \in \mathbb{R}^{1 \times N}$;
6. Rearrange $\mathbf{p}$ in ascending order from the smallest to the largest value.
7. Extract the $\mathcal{N}$ smallest values and find the corresponding images.
8. **return** $\mathcal{N}$ most similar images.

end

Figure 3.2: The CBIR-BoSE algorithm.

3.2 Performing CBIR with BoSE

The high-dimensional feature data extracted from each of the datasets is reduced to a fewer number of dimensions via BoSE, the intent being to perform retrieval in the BoSE reduced space. Briefly, the retrieval is performed as follows. The query sample and all existing annotated database samples are aggregated and the BoSE representation for all images (following feature extraction and weighting) is determined. Using the EDM, the distance between the query image and all of the database images is calculated in the BoSE space. The resulting distance vector is sorted in ascending order and the most similar database images in terms of distance are outputted. The CBIR-BoSE algorithm is illustrated in Figure 3.2.

Chapter 4

Experimental Design and Evaluation

4.1 Dataset Description

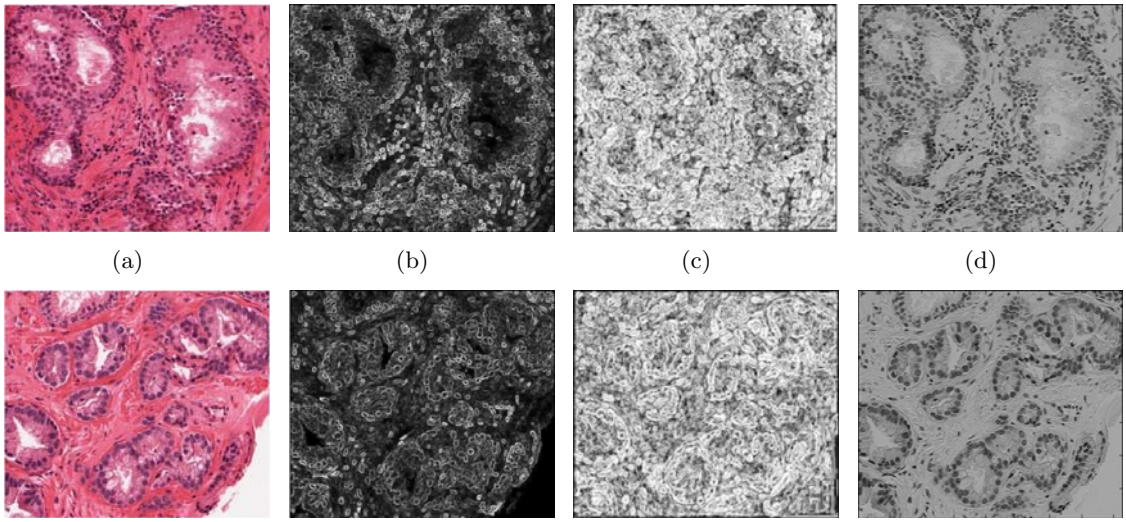
We considered three datasets (Table 4.1). Slides from all three datasets were stained with hematoxylin and eosin (H&E) and scanned into a computer via a whole-slide digital scanner at the University of Pennsylvania (prostate cancer) and the Cancer Institute of New Jersey (breast cancer). The prostate and breast cancer images were taken at magnifications of 40x and 20x, respectively, and were saved in the SVS format. Pathologists were instructed to manually place a contour around homogeneous regions of tissue corresponding to either "cancer" or "non-cancer" regions. Annotation was performed on the scanned SVS biopsy image files using the ImageScope software platform (Aperio, Inc.). No confounding tissue types (e.g. atrophy, prostatic intraepithelial neoplasia) were included. The entire tissue biopsy images were then divided into 30-by-30 square pixel regions; within these 900 pixels, if over 50% of the pixels (450) contained the expert's annotation, those regions were included in the dataset. All of the images were converted from the RGB color space to the HSV (Hue, Saturation, Value) space to mitigate the effect of varying stain intensities. By converting images to the HSV space, we ensure that any potential stain intensity variation across images is confined to a single channel (the "Value" channel). The objective of Experiment 1 ($\mathcal{D}_1$) was to distinguish between malignant and benign prostate tissue patches (Table 4.1, Figure 4.1) from biopsy samples obtained from 58 patients. In Experiment 2 ($\mathcal{D}_2$), we aimed to distinguish betwen high and low grade breast cancer tissue patches from biposy samples obtained from 55 patients. Lastly, the objective of Experiment 3 ($\mathcal{D}_3$) was to distinguish between high and low levels of LI in breast cancer tissue patches from biopsy samples obtained from 12 patients. For each of the experiments we compared the CBIR-BoSE and CBIR-SE systems in terms of CBIR-based and classifier-based performance measures.

Data	**Classes** $(+1/-1)$	**Class Distribution** $(+1/-1)$	**No. of Samples**
Prostate ($\mathcal{D}_1$)	Cancer/Benign	29/29	58
Breast ($\mathcal{D}_2$)	High Grade/Low Grade	36/19	55
Breast ($\mathcal{D}_3$)	High LI/Low LI	20/21	41

Table 4.1: List of the breast cancer and prostate cancer datasets used in this study.

4.2 Experiment 1: Distinguishing Malignant from Benign Prostate Histopathology

Hematoxylin stains the nuclei and gives it shade of purple, while eosin stains the cytoplasm and stroma and gives them a shade of pink. Benign prostate histology exhibits amorphous glands with epithelial cells lining the glands. They display a lighter color when stained with H&E because they have a lower amount of nuclear proliferation. Malignant prostate histology displays more ordered glands that are lined with epithelial cells containing more nucleoli within the nuclei. The greater number of nucleoli cause the cells to display a darker stain when compared to a benign prostate tissue sample. Doyle, et al. [33] showed that Gabor, Haralick, and first-order statistics can discriminate well between benign and malignant regions of prostate tissue. The Gabor filter is a Gaussian function that is modulated by a sinusoid. Gabor features quantify the response of image regions with intensity patterns that are similar to the Gabor filter's orientation and frequency parameters [34][35]. Haralick features are able to capture the underlying patterns in the image texture through the construction of a co-occurrence matrix [36]. Co-occurrence image features are based on the adjacency of pixel values in an image. An adjacency matrix is created where the value of the i-th row and the j-th column equals the number of times pixel values i and j appear within a fixed distance of one another. First-order statistical features quantify the intensity variations in the image. These texture features are related to the degree of nuclear proliferation, where there are more gradients as a result of the transitions between high intensity values (the stroma/lumen) and the low intensity values (nuclei and nucleoli). In [33], 14 highly discriminating pixel-wise features were learned via AdaBoost [23] out of a feature set that comprised over 900 features. AdaBoost assigned a weight to all of the features and these weights were thresholded in that features with $\alpha > 0.05$ were retained while the other features were discarded. In the current study these 14 features were extracted for



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

(e) (f) (g) (h)

Figure 4.1: Examples of (a) benign and (e) gleason grade 3 prostate cancer images and their corresponding feature images: (b) (f) first-order statistics (Range using a 5×5 window, Hue color channel) , (c) (g) Haralick (Correlation using a 5×5 window, Hue color channel), and (d) (h) Gabor features (5×5 window, $\theta = \frac{\pi}{6}$, Hue color channel).

each image, generating 14 corresponding feature images. The pixel values for each feature image were averaged, generating a 14 element feature vector to characterize each prostate image (Table 4.2). Figure 4.1 displays the texture feature images.

Texture Feature	**Parameters**
First-order Statistics (Standard Deviation, Range)	Window size: $w = 5$
Haralick Features (Information Measure, Correlation, Energy, Contrast Variance, Entropy)	Window size: $w = 5$ Distance: $\delta = 1$
Gabor Features	Window size: $w \in \{5, 9\}$ Orientation: $\theta \in \{0, \frac{\pi}{6}, \ldots, \frac{5\pi}{6}\}$

Table 4.2: Texture features extracted from the prostate tissue sample images.

4.3 Experiment 2: Distinguishing High from Low Grade Breast Histopathology

Two of the defining histological features of breast cancer are the disorganization of the tissue and the structure of the cells. The severity of the cancer is given a Bloom Richardson (BR) grade level [27]. Breast cancer tissue samples with greater disorganization and increasingly irregular structure are given higher grades. High grade samples exhibit more nuclear proliferation than low grade samples. As with the prostate cancer samples, the breast cancer biopsy samples were stained with H&E. Haralick features were extracted and used to describe the degree of nuclear proliferation by quantifying the variations in the intensity values in the images. The objective of this experiment was to retrieve images corresponding to the grade of the query image. To define a two-class problem, all images are first separated into either low (BR 4, 5) and high (BR 7, 8) grade classes (Table 4.1). From each image, 12 Haralick feature images were generated and the following statistics were computed from the pixel values from each feature image: mean, standard deviation, and entropy. This was done for all three color channels in the HSV space.

4.4 Experiment 3: Distinguishing High LI from Low LI Breast Histopathology

The class problem is defined as follows: images were separated into either low LI or high LI classes (Table 4.1). To quantify the arrangement of lymphocytic nuclei in the histology images, architectural features were computed for each image. The centroids of the lymphocytic nuclei are used to construct the Delaunay Triangulation $\mathcal{G}_D$ (Figures 4.2 (b) and 4.2 (f)), the Minimum Spanning Tree $\mathcal{G}_M$ (Figures 4.2 (c) and 4.2 (g)), and the Voronoi Diagram $\mathcal{G}_V$ (Figures 4.2 (d) and 4.2 (h)). Automated nuclear detection was performed to identify the nuclear centers as centroids of the different graphs. However, the cancer and lymphocytic nuclei are similar in appearance. In general, lymphocytic nuclei differ in appearance from cancer cell nuclei by their smaller size, more circular shape, and a darker homogeneous staining [37]. We took these differences into account and performed automated nuclear detection in the following manner.

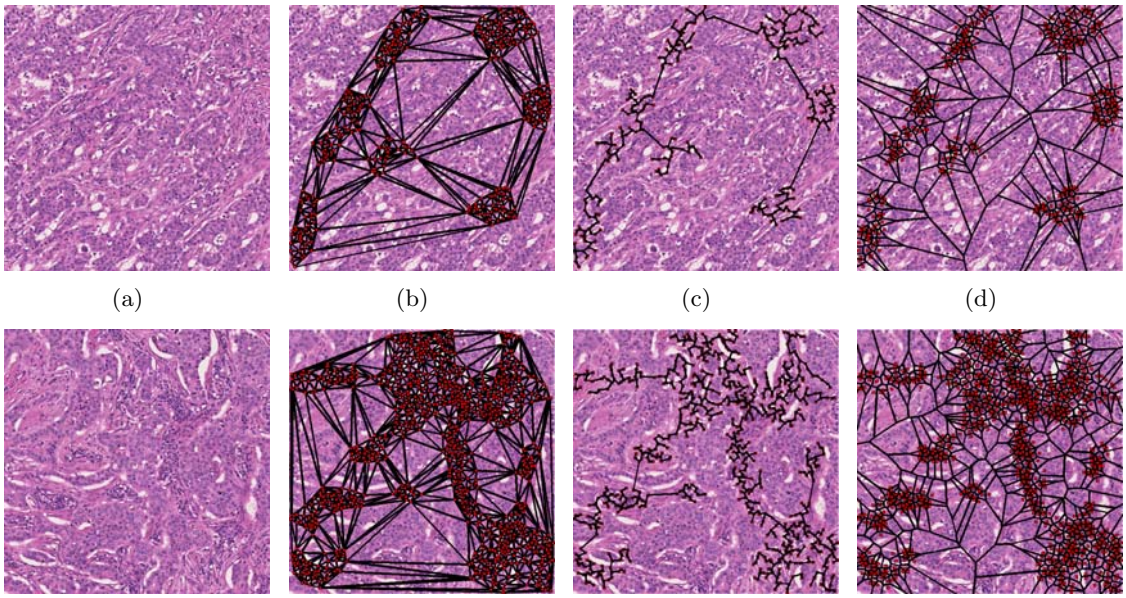


Figure 4.2: Example breast histopathology images that contain (a) low and (e) high levels of lymphocytic infiltration with their corresponding feature images: (b) (f) delaunay triangulation, (c) (g) minimum spanning tree, and (d) (h) voronoi graphs. Quantitative graph features were calculated using the graphs constructed on the image.

Step 1: On each image, M candidate nuclear centers $M = \{m_1, m_2, \ldots, m_M\}$ were found by convolving the image $\mathbf{x}_i$ with a Gaussian (smoothing) kernel at multiple scales. This was done to account for the variation in lymphocyte size. The darkest pixels were found on the smoothed image based on local differences in luminance and these were the candidate lymphocytic nuclear centers.

Step 2: Using the region-growing scheme [38], each of the M candidate lymphocytic nuclear centers was grown into a corresponding region $\mathbf{R}$. The optimal regions were identified when the boundary strength, which is defined as the difference in the mean intensity of the pixels in the internal boundary and the current boundary of the region, was at a maximum. See [37] for a more detailed description.

Step 3: Each of $r \in \mathbf{R}$ contained two random variables: $A_r \in \{\omega_c, \omega_l\}$ which is the classification of the candidate nuclear centers as either a cancer (ω_c) or lymphocytic (ω_l) nucleus and $B_r \equiv [C_r, \phi_r]^{\mathrm{T}} \in \mathbb{R}^{+2}$ where C_r is the square root of the nuclear area and ϕ_r is the standard deviation of the luminace in the nuclear region. The labels,

Graph	Features
Voronoi Diagram (13 features)	Total area of all polygons Polygon Area (mean, standard deviation, min/max ratio, entropy) Polygon Perimeter (mean, standard deviation, min/max ratio, entropy) Polygon Chord Length (mean, standard deviation, min/max ratio, entropy)
Delaunay Triangulation (8 features)	Triangle Area (mean, standard deviation, min/max ratio, entropy) Triangle Side Length (mean, standard deviation, min/max ratio, entropy)
Minimum Spanning Tree (4 features)	Branch Length (mean, standard deviation, min/max ratio, entropy)
Nuclear Features (25 features)	Density of nuclei Distance to $\{3, 5, 7\}$ nearest nuclei (mean, standard deviation, disorder) Number of nuclei in a $\{10, 20, \ldots, 50\}$ pixel radius (mean, standard deviation, disorder)

Table 4.3: List of the features extracted to quantify the degree of LI. A detailed description of the feature extraction and graph construction can be found in [1].

A_r, given the feature vectors B_r are estimated via a maximum *a posteriori* (MAP) estimation by finding the A_r that maximizes the posterior probability

$$p(A_r|B_r) = \frac{p(B_r|A_r)p(A_r)}{p(B_r)} \tag{4.1}$$

where $p(B_r|A_r)$ is the liklihood term and $p(A_r)$ and $p(B_r)$ are prior distributions. $p(B_r)$ is ignored because maximization was done with respect to $p(A_r)$.

Step 4: $p(B_r|A_r)$ is computed from PDFs, where A_r is provided by manual delineation of lymphocytes in a training set.

Step 5: The prior distribution $p(A_r)$ is defined by a Markov Random Field (MRF) and computed. The iterated conditional modes (ICMs) algorithm [39], a derterministic relaxation procedure, was used to compute the MAP estimation and classify each $r \in \mathbf{R}$. The regions classified as cancer nuclei were discarded and the centriods of the lymphocytic nuclei were calculated, yielding $O = \{o_1, o_2, \ldots, o_L\}$ where $O \subseteq M$. Details of the automated nuclear detection can be found in [37].

Using the O centroids, we constructed a graph $\mathcal{G} = (\mathbf{V}, \mathbf{E}, \mathbf{J})$, where $\mathbf{V}$ represent the vertices of the graph which correspond to the number of centroids, $\mathbf{E}$ are the set of edges, and $\mathbf{J}$ are the weights of the edges, proportional to edge length. The set of vertices, edges, and weights make up a unique graph on the image. From each graph, we extracted a set of features listed in Table 4.3. A detailed description of the graph construction and feature extraction can be found in [1].

4.5 Evaluation Measures

4.5.1 CBIR-BoSE

The performance of a CBIR system is determined by how many retrieved images for a given query image are relevant to the query, defined as images which belong to the same class as the query image, and also the order in which they appear. Precision is defined as $\rho(\beta) = \frac{\xi(\beta)}{\beta}$, where $\xi(\beta)$ denotes the number of relevant objects in the β closest objects. Recall is defined as $r(\beta) = \frac{\xi(\beta)}{\xi(N-1)}$. Precision-recall curves were generated by plotting $\rho(\beta)$ versus $r(\beta)$ for $\beta \in \{1, 2, \ldots, N-1\}$. Area under the precision-recall curve (AUPRC) was measured and used to evaluate the CBIR system. The AUPRC values $\in [0, 1]$ where an AUPRC of 1 indicates that the CBIR system only retrieved relevant images and an AUPRC of 0 indicates that the CBIR system only retrieved irrelevant images. Therefore, the higher the AUPRC, the better the CBIR system. We denote θ^{AU}_{BoSE} and θ^{AU}_{SE} as the AUPRC values for CBIR-BoSE and CBIR-SE, respectively.

4.5.2 Classifier Evaluation of BoSE and SE

A second performance measure for evaluating BoSE is classifier accuracy. Of the classifiers available (Support Vector Machines, Neural Nets, etc.), the Random Forest (RF) classifier was chosen due to its ability to accurately and efficiently run on large databases with minimal training time and lower overall computational time. The RF classifier (obtained by bagging decision trees) [40] is trained on both $\mathcal{M}^{\text{BoSE}}$ and $\mathcal{M}^{\text{SE}}$ (Figure 5.3). The accuracy of the RF classifier should reflect the class discriminability of $\mathcal{M}^{\text{BoSE}}$. A RF classifier is an ensemble of decision trees (i.e. weak learners) combined via bootstrap aggregation.

Averaging decisions across weak learners creates a strong learner that reduces overall bias and variance [40]. We define θ_{BoSE}^{Acc} amd θ_{SE}^{Acc} as the classification accuracy when performing classification in the lower dimensional spaces created by BoSE and SE, respectively. The classification accuracy is defined as $\frac{TP+TN}{TP+TN+FP+FN}$ where TP are the true positives, TN are the true negatives, FP are the false positives, and FN are the false negatives.

Let $S_{+1} \subset \mathbf{X}$ and $S_{-1} \subset \mathbf{X}$ where for any $\mathbf{a} \in S_{+1}$, $\mathcal{L}(\mathbf{a}) = +1$ and for any $\mathbf{b} \in S_{-1}$, $\mathcal{L}(\mathbf{b}) = -1$. S_{+1} and S_{-1} are subsets of the total number of the specific class objects we have in $\mathbf{X}$. S_{+1} and S_{-1} are randomly sampled with replacement from $\mathbf{X}$, ensuring that each of S_{+1} and S_{-1} only comprise of instances from either of $+1$ and -1. Each random sampling of S_{+1} and $S_{-1}$is used to train a decision tree classifier Ω_v, where $v \in \{1, 2, \ldots, V\}$ and so that $\Omega_v(\mathbf{x}) \in \{+1, -1\}$.

Randomized, 3-fold cross-validation was used to determine training and testing inputs for the RF classifier. First, the entire dataset $\mathbf{X}$ was randomly divided into three equally-sized subsets $\mathbf{X}^1, \mathbf{X}^2, \mathbf{X}^3 \subset \mathbf{X}$. Two of the subsets were used for training the RF classifier, which was then evaluated on the remaining subset. The subsets were subsequently rotated until each subset was used for evaluation exactly once. The entire cross-validation scheme was repeated over 50 iterations, over which the mean and standard deviation of the classification accuracy were reported.

4.5.3 Evaluating Intrinsic Dimensionality for CBIR-BoSE

When performing retrieval and classification in the lower-dimensional space, identifying the optimal number of dimensions within which to embed the data is a non-trivial task. Each dataset possesses an intrinsic dimensionality in which the classification accuracy and the retrieval performance will be optimal. In order to evaluate the effect of the total number of embedding dimensions to be considered, for the purpose of maximizing classification accuracy and the AUPRC, each dataset was reduced to various lower dimensional embeddings. The corresponding number of dimensions associated with these reduced dimensional embeddings was varied and BoSE was evaluated in these different spaces (Table 4.4). We define $\theta_{BoSE,k}^{Acc}$ and $\theta_{BoSE,k}^{AU}$ as the accuracy and AUPRC using BoSE in k dimensions, where $k \in \{1, 2, \ldots, K\}$ and similarly $\theta_{SE,k}^{Acc}$ and $\theta_{SE,k}^{AU}$ for SE. The maximum, minimum,

and average AUPRC and classification accuracy is reported and calculated in the following manner: $\theta_{\nu}^{\mu,max} = \max_{k}[\theta_{\nu,k}^{\mu}]$, $\theta_{\nu}^{\mu,min} = \min_{k}[\theta_{\nu,k}^{\mu}]$, $\psi_{\nu}^{\mu} = \frac{1}{K}\sum_{k=1}^{K}\theta_{\nu,k}^{\mu}$ where $\mu \in \{Acc, AU\}$ and $\nu \in \{BoSE, SE\}$.

Dataset	**Original Dimensionality**	**Reduced Dimensionality**
Prostate Cancer	14	2, 3, 4, 5, 6, 7
Breast Cancer Grading	108	2, 3, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50
Lymphocytic Infiltration	50	2, 3, 5, 10, 15, 20, 25

Table 4.4: The original dimensionality of the data and its reduced dimensionality employed for evaluating CBIR-BoSE and CBIR-SE. Both CBIR systems were evaluated after projecting the original high dimensional data into spaces of progressively different reduced dimensions.

Chapter 5

Results and Discussion

5.1 Experiment 1: Distinguishing Malignant from Benign Prostate Histopathology

5.1.1 Quantitative Evaluation

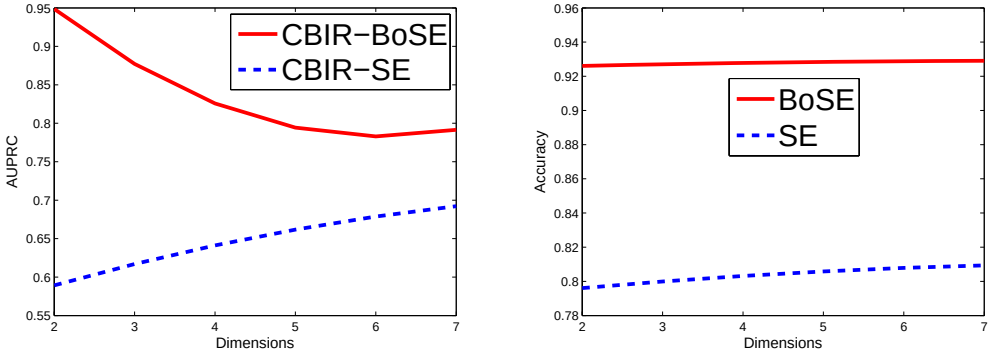
Figure 5.1 and Table 5.2 reveal that over a range of dimensions, CBIR-BoSE consistently outperforms CBIR-SE in terms of (a) AUPRC, and (b) accuracy. It appears that retrieval with BoSE seems to improve with lower embedding dimensionality. For $\mathcal{D}_1$, $\theta_{BoSE}^{AU,max}$ and $\theta_{BoSE}^{AU,min}$ were greater than $\theta_{SE}^{AU,max}$ and $\theta_{SE}^{AU,min}$ (Table 5.1). The average AUPRC for CBIR-BoSE (ψ_{BoSE}^{AU}) across the all the dimensionalities evaluated was greater than the average AUPRC for CBIR-SE (ψ_{SE}^{AU}) (Table 5.1). $\theta_{BoSE}^{Acc,max}$ and $\theta_{BoSE}^{Acc,min}$ were greater than $\theta_{SE}^{Acc,max}$ and $\theta_{SE}^{Acc,min}$ (Table 5.2). ψ_{BoSE}^{Acc} was greater than ψ_{SE}^{Acc} and is statistically significant using a $p < 0.05$. Unlike the AUPRC values, the accuracy values remain relatively invariant to the number of dimensions that $\mathcal{D}_1$ is embedded into via BoSE and SE.

Dataset	$\theta_{BoSE}^{AU,max}$	$\theta_{SE}^{AU,max}$	$\theta_{BoSE}^{AU,min}$	$\theta_{SE}^{AU,min}$	ψ_{BoSE}^{AU}	ψ_{SE}^{AU}
$\mathcal{D}_1$	**0.87**	0.66	**0.70**	0.60	**0.77**	0.63
$\mathcal{D}_2$	**0.90**	0.90	**0.74**	0.57	**0.80**	0.68
$\mathcal{D}_3$	**0.59**	0.43	**0.45**	0.36	**0.51**	0.39

Table 5.1: Quantitative results showing the maximum, minimum, and mean AUPRC values for Experiment 1 ($\mathcal{D}_1$), Experiment 2 ($\mathcal{D}_2$), and Experiment 3 ($\mathcal{D}_3$). ψ_{BoSE}^{AU} is greater than ψ_{SE}^{AU} for $\mathcal{D}_1$, $\mathcal{D}_2$, and $\mathcal{D}_3$ and is statistically significant using a $p < 0.05$.

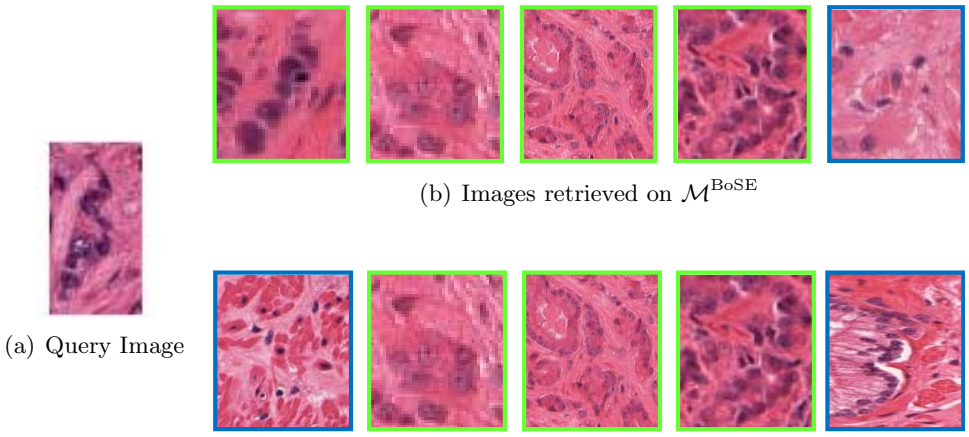
Dataset	$\theta_{BoSE}^{Acc,max}$	$\theta_{SE}^{Acc,max}$	$\theta_{BoSE}^{Acc,min}$	$\theta_{SE}^{Acc,min}$	ψ_{BoSE}^{Acc}	ψ_{SE}^{Acc}
$\mathcal{D}_1$	**0.93**	0.81	**0.92**	0.80	**0.93**	0.80
$\mathcal{D}_2$	**0.99**	0.99	**0.94**	0.93	**0.98**	0.98
$\mathcal{D}_3$	**0.96**	0.92	**0.90**	0.90	**0.94**	0.91

Table 5.2: Quantitative results showing the maximum, minimum, and mean classification accuracies for Experiment 1 ($\mathcal{D}_1$), Experiment 2 ($\mathcal{D}_2$), and Experiment 3 ($\mathcal{D}_3$).



(a) (b)

Figure 5.1: Quantitative results displaying (a)$\theta_{BoSE,k}^{AU}$, $\theta_{SE,k}^{AU}$ and (b) $\theta_{BoSE,k}^{Acc}$, $\theta_{SE,k}^{Acc}$ over the dimensions $k \in \{2, \ldots, 7\}$ for Experiment 1. A second order polynomial was fitted to the data to illustrate the trends in θ^{AU} and θ^{Acc}.



(a) Query Image

(b) Images retrieved on $\mathcal{M}^{\text{BoSE}}$

(c) Images retrieved on $\mathcal{M}^{\text{SE}}$

Figure 5.2: The illustration shows the retrieved images using (b) BoSE and (c) SE for (a) the query image (prostate cancer tissue sample). The images that are outlined in green and blue are from the cancer and benign classes, respectively. For the top five retrieved images, CBIR-BoSE returned more relevant images compared to CBIR-SE.

5.1.2 Qualitative Evaluation

For each of the top five images retrieved, CBIR-BoSE yielded more relevant images compared to CBIR-SE (Figure 5.2) reflecting that objects from the same class are mapped closer to each other in $\mathcal{M}^{\text{BoSE}}$. Figures 5.3(a) and 5.3(d) display $\mathcal{M}^{\text{BoSE}}$ and $\mathcal{M}^{\text{SE}}$, respectively, showing a much greater separation between the malignant and benign classes in $\mathcal{M}^{\text{BoSE}}$ compared to $\mathcal{M}^{\text{SE}}$.

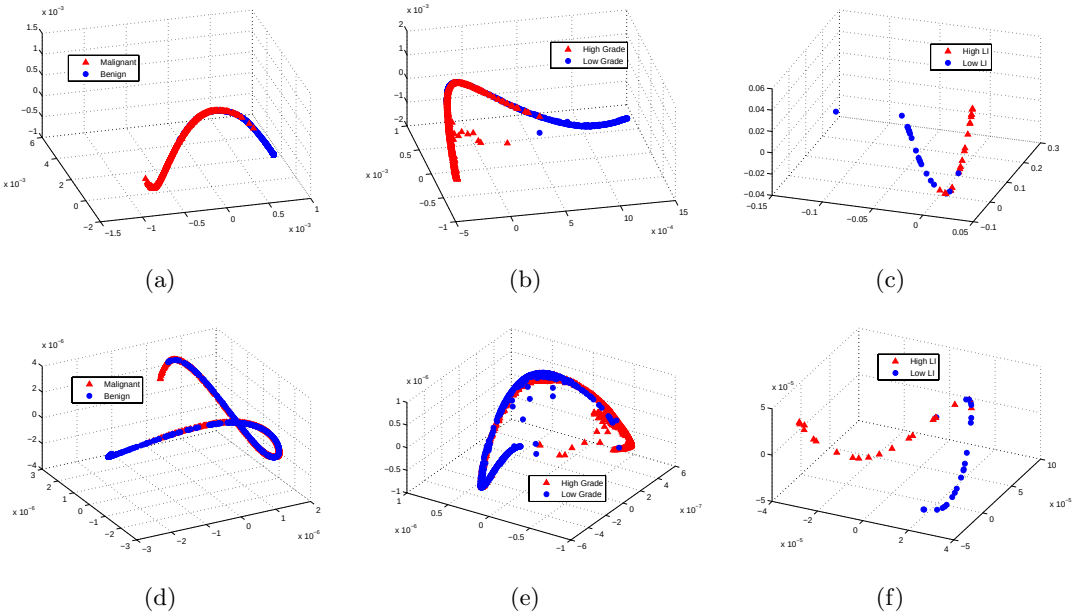


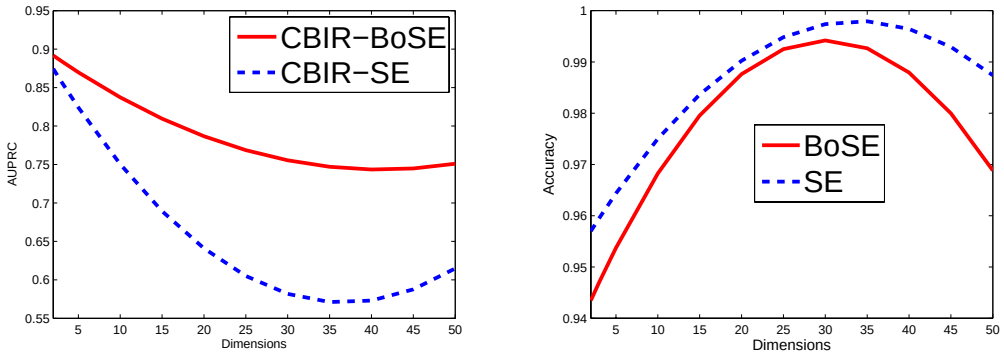
Figure 5.3: $\mathcal{M}^{\text{BoSE}}$ and $\mathcal{M}^{\text{SE}}$ shown for (a), (d) $\mathcal{D}_1$, (b), (e) $\mathcal{D}_2$, and (c), (f) $\mathcal{D}_3$ using (a), (b), (c) BoSE and (d), (e), (f) SE. Although the low-dimensional data does not appear as a set of 'clusters', we can see a clear class separation on the manifold when using BoSE (top row) compared to SE (bottom row).

5.2 Experiment 2: Distinguishing High from Low Grade Breast Histopathology

5.2.1 Quantitative Evaluation

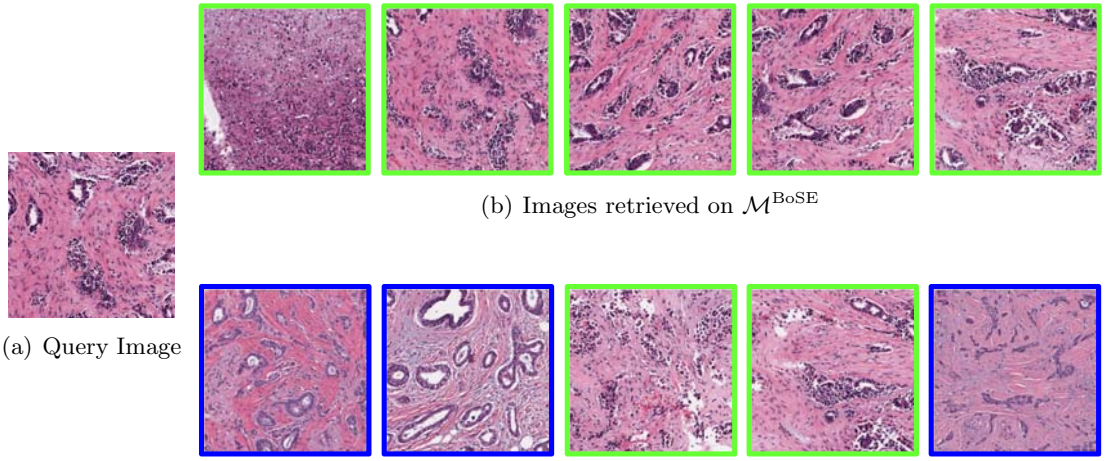
For $\mathcal{D}_2$, θ^{AU}_{BoSE} and θ^{AU}_{SE} decreased as the dimensionality of the data increased (Figure 5.4). While $\theta^{AU,max}_{BoSE}$ and $\theta^{AU,max}_{SE}$ occured when $\mathcal{D}_2$ was reduced to two dimensions and were similar, θ^{AU}_{SE} decreased more drastically compared to θ^{AU}_{BoSE}. This resulted in $\theta^{AU,min}_{BoSE}$ being greater than $\theta^{AU,min}_{SE}$ (Table 5.1). Another consequence of the difference in the rate of

decrease of θ^{AU} between CBIR-BoSE and CBIR-SE was that ψ^{AU}_{BoSE} was greater compared to ψ^{AU}_{SE} (Table 5.1). $\theta^{Acc,max}$, $\theta^{Acc,min}$, and ψ^{Acc} yielded similar values for both BoSE and SE and no appreciable difference was observed (Table 5.2).



(a) (b)

Figure 5.4: Quantitative results displaying (a) $\theta^{AU}_{BoSE,k}$, $\theta^{AU}_{SE,k}$ and (b) $\theta^{Acc}_{BoSE,k}$, $\theta^{Acc}_{SE,k}$ over all the dimensions $k \in \{2, 3, 5, 10, \ldots, 50\}$ for the breast cancer images. θ^{AU}_{BoSE} is greater than θ^{AU}_{SE}. A second order polynomial was fitted to the data to illustrate the trends in θ^{AU} and θ^{Acc}.



(a) Query Image

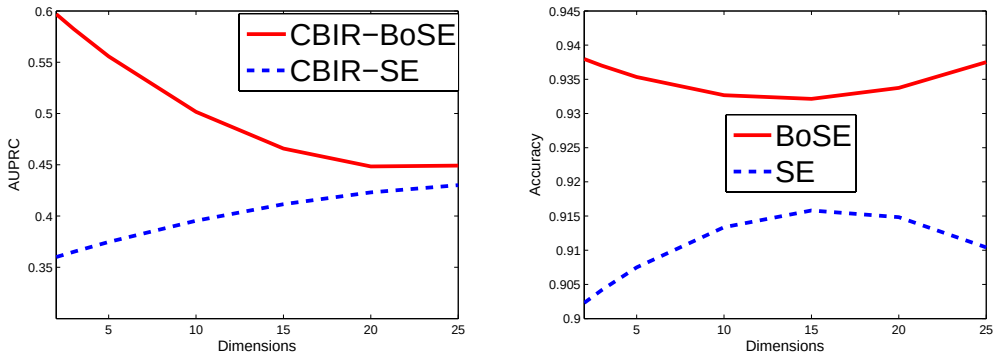
(b) Images retrieved on $\mathcal{M}^{\text{BoSE}}$

(c) Images retrieved on $\mathcal{M}^{\text{SE}}$

Figure 5.5: The illustration shows the retrieved images using (b) BoSE and (c) SE for (a) the query image (high grade breast cancer tissue sample). The images that are outlined in green and blue are from the high and low grade breast cancer classes, respectively. For the top five retrieved images, CBIR-BoSE returned more relevant images compared to CBIR-SE.

5.2.2 Qualitative Evaluation

Figure 5.5 displays the top five images for both the CBIR-BoSE and CBIR-SE systems. CBIR-BoSE retrieved more relevant images and thus illustrated that images from similar classes are mapped closer to each other in $\mathcal{M}^{\text{BoSE}}$ compared to $\mathcal{M}^{\text{SE}}$. $\mathcal{M}^{\text{BoSE}}$ (Figure 5.3 (b)) appears to suggest better class separability compared to SE (Figure 5.3 (e)).



(a) (b)

Figure 5.6: Quantitative results displaying (a) $\theta^{AU}_{BoSE,k}$, $\theta^{AU}_{SE,k}$ and (b) $\theta^{Acc}_{BoSE,k}$, $\theta^{Acc}_{SE,k}$ over all the dimensions $k \in \{2, 3, 5, 10, \ldots, 25\}$ for the lymphocytic infiltration images. θ^{AU} and θ^{Acc} for BoSE were greater compared to SE. A second order polynomial was fitted to the data to illustrate the trends in θ^{AU} and θ^{Acc}.

5.3 Experiment 3: Distinguishing High LI from Low LI Breast Histopathology

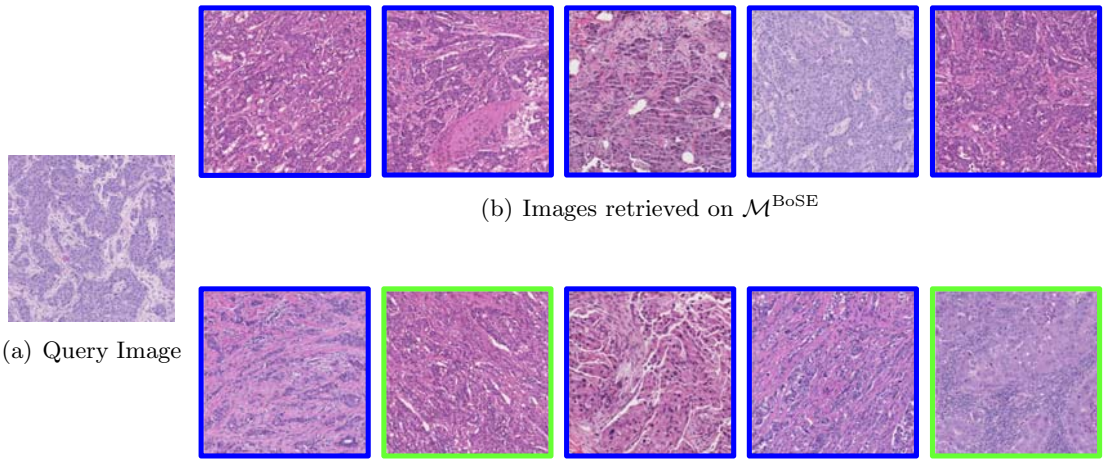
5.3.1 Quantitative Evaluation

For $\mathcal{D}_3$, $\theta^{AU,max}_{BoSE}$, $\theta^{AU,min}_{BoSE}$, and ψ^{AU}_{BoSE} were greater compared to $\theta^{AU,max}_{SE}$, $\theta^{AU,min}_{SE}$, and ψ^{AU}_{SE} (Figure 5.6, Table 5.1). $\theta^{Acc,max}_{BoSE}$ and ψ^{Acc}_{BoSE} were higher compared to $\theta^{Acc,max}_{SE}$ and ψ^{Acc}_{SE} (statistically significant with $p < 0.05$), but $\theta^{Acc,min}_{BoSE}$ was similar to $\theta^{Acc,min}_{SE}$ (Table 5.2). The dimensionality of the data had little effect on the θ^{Acc}_{BoSE} and θ^{Acc}_{SE}.

5.3.2 Qualitative Evaluation

Figure 5.7 displays the top five images for both the CBIR-BoSE and CBIR-SE systems. $\mathcal{M}^{\text{BoSE}}$ (Figure 5.3 (c)) appears to show better separation between the images that have

low LI and images that have high LI than $\mathcal{M}^{\text{SE}}$ (Figure 5.3 (f)).

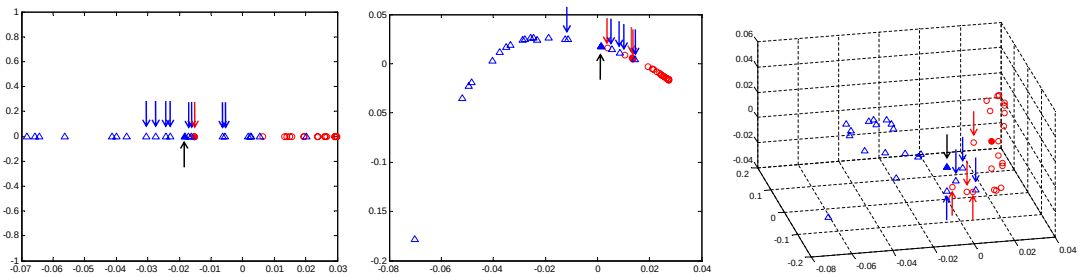


(a) Query Image

(b) Images retrieved on $\mathcal{M}^{\text{BoSE}}$

(c) Images retrieved on $\mathcal{M}^{\text{SE}}$

Figure 5.7: The illustration shows the retrieved images using (b) BoSE and (c) SE for (a) the query image (low LI breast cancer tissue sample). The images that are outlined in green and blue are from the high LI and low LI classes, respectively. In the top five retrieved images, CBIR-BoSE returned more relevant images compared to CBIR-SE.



(a) (b) (c)

Figure 5.8: The LI data embedded into $\mathcal{M}^{\text{BoSE}}$ in (a) $\mathbb{R}^1$, (b) $\mathbb{R}^2$, and (c) $\mathbb{R}^3$. The filled in blue triangle denotes the query image and the arrows denote its eight nearest images. When the dimensionality of $\mathcal{M}^{\text{BoSE}}$ is low, most of the eight nearest images are from the same class as the query image. However, as the dimensions are increased more irrelevant images are part of the query image's eight nearest neighbors. Hence, the AUPRC decreases as the number of dimensions is increased.

5.4 AUPRC as a Function of Increasing Dimensionality of $\mathcal{M}^{\text{BoSE}}$

θ^{AU}_{BoSE} decreased as the dimensionality of $\mathcal{M}^{\text{BoSE}}$ increased for all three experiments. We offer some intuition as to why this happens. Let the blue triangle in Figure 5.8 denote the query image. When the dataset is embedded into a one dimensional space, seven of the eight nearest samples are from the same class. Thus, when performing image retrieval, the majority of the top eight retrieved images will be relevant. When the data is embedded into a two dimensional space, only four of the eight nearest images are from the same class. If image retrieval is performed in this space, only half of the top eight images retrieved will be relevant, reducing precision for that query image; however, classification accuracy for the whole dataset is unchanged. Lastly, when the data is embedded into a three dimensional space, a similar situation is encountered. It should be noted that because classification and training is performed each time a dataset is reduced in dimensionality, it is very possible that all of these spaces will yield either similar classification accuracies or improvements in classification accuracy. Consequently, the apparent discrepancy between the trends in AUPRC and accuracy for BoSE and SE across a different number of dimensions exists because in CBIR the *order* of the retrieved data points affects the AUPRC while the accuracy is unaffected.

Chapter 6

Concluding Remarks and Future Work

In this work, we presented a CBIR system that utilized Boosted Spectral Embedding (BoSE), which employed the boosted distance metric (BDM) in conjunction with Spectral Embedding (SE). The BDM preferentially weights features that discriminate between objects of different classes allowing for a similarity matrix which better describes object similarity. We have created a task-specific embedding technique that improves class separability, yielding better classification and retrieval. In this work we applied the CBIR-BoSE framework in the context of problems in digital pathology. SE has been shown to be less sensitive to the choice of system parameters compared to other popular manifold learning schemes (e.g. Isomaps [13], LLE [14]). The CBIR system presented here could be employed as a teaching tool for pathology residents and fellows. Specifically, we focused on distinguishing between (1) benign and malignant prostate histology, (2) low and high grade ER+ breast cancer histology, and (3) low and high levels of lymphocytic infiltration (LI) in HER2+ breast tissue. We compared CBIR-BoSE to CBIR-SE, which uses the Euclidean distance metric to define object similarity. For different numbers of dimensions of the low dimensional space, for different datasets, for different performance measures (CBIR and classifier based), CBIR-BoSE outperformed CBIR-SE a majority of the time.

One of the current limitations of our CBIR system is that for every new query image, the manifold for the query along with all existing database images needs to be computed. This procedure needs to be repeated for each new query instance. In the future work we are looking to incorporate out of sample extrapolation schemes [41] which allow for the mapping of a new query instance into an existing lower dimensional space, without having to recompute the eigenvalue decomposition; thus reducing the overall computational cost of a new retrieval task. We also intend to extend our current scheme to the multi-class case.

The CBIR-BoSE system consists of two computational portions: the online and the offline portions. The offline portion consists of the computation of the PDFs for the weak classifiers, which is the most computationally expensive step. The amount of time it takes to compute the PDFs is dependent solely on the number of features used to represent an image in a particular dataset. For example, the prostate dataset was comprised of 14 features and it took approximately eight minutes to compute the PDFs. However, this step can be performed offline because the PDFs are not recomputed for each new query image. The online portion consists of the computation of the weak classifier weights, the low dimensional manifold using BoSE, and the distance between the query image and the database images. The run-time for the weight and the query image-database image distance computations are negligible, but the run-time for the creation of a low dimensional manifold that incorporates the query image is dependent on the number of samples. For example, the LI dataset consisted of 41 patients (samples) and it took approximately one minute to compute the low dimensional manifold via BoSE. For use in a clinical setting, the system can compute the PDFs when it is not being used by physicians so as to not cause any delay in image retrieval. A graphical user interface needs to be created for the ease of use for the physicians. All experiments performed in this thesis were run on a Linux (Ubuntu 2.6.22-15-generic) server with the following specifications: 2.33 GHz clock speed, 32 gigabytes of random access memory, a Super Micro X7DBE+ mother board, and a 2x quad-core Xeon E5345 processor.

References

[1] S. Doyle, M. Hwang, K. Shah, A. Madabhushi, M. Feldman, and J. Tomaszeweski. Automated grading of prostate cancer using architectural and textural image features. In *International Symposium on Biomedical Imaging*, pages 1284 –1287, 2007.

[2] S. Doyle, M. Hwang, S. Naik, M. Feldman, J. Tomaszeweski, and A. Madabhushi. Using manifold learning for content-based image retrieval of prostate histopathology. In *Medical Image Computing and Computer Assisted Intervention*, pages 53–62, 2007.

[3] C.K. Reddy and F.A. Bhuyan. Retrieval and ranking of biomedical images using boosted haar features. In *International Conference on BioInformatics and BioEngineering*, pages 1–6, 2008.

[4] L. Yang, R. Jin, L. Mummert, A. Suthankar, B. Goode, B. Zheng, S. Hoi, and M. Satyanarayanan. A boosting framework for visuality-preserving distance metric learning and its application to medical image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32:30–44, 2010.

[5] J. G. Dy, C. E. Brodley, A. Kak, L. S. Broderick, and A. M. Aisen. Unsupervised feature selection applied to content-based retrieval of lung images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25:373–378, 2003.

[6] K. Bunte, N. Petkov, M. Biehl, and F. Jonkman. Adaptive metrics for content based image retrieval in dermatology. In *European Symposium on Aritificial Neural Networks*, pages 129–134, 2009.

[7] N. Mehta, R.S. Alomari, and V. Chaudhary. Content based sub-image retrieval system for high resolution pathology images using salient interest points. In *Engineering in Medicine and Biology Society*, pages 3719 –3722, 2009.

[8] J. C. Caicedo, F. A. Gonzalez, and E. Romero. A semantic content-based retrieval method for histopathology images. In *Proceedings of the 4th Asia Information Retrieval Conference on Information Retrieval Technology*, AIRS'08, pages 51–60, Berlin, Heidelberg, 2008. Springer-Verlag.

[9] X. He, D. Cai, and J. Han. Learning a maximum margin subspace for image retrieval. *IEEE Transactions on Knowledge and Data Engineering*, 20(2):189–201, February 2008.

[10] J.H. Huang, A. Zia, J. Zhou, and A. Robles-Kelly. Content-based image retrieval via subspace-projected salient features. *Digital Image Computing: Techniques and Applications*, 0:593–599, 2008.

[11] I. Joliffe. *Principle Component Analysis*. Springer-Verlag, 1986.

[12] G. Lee, C. Rodriguez, and A. Madabhushi. Investigating the efficacy of nonlinear dimensionality reduction schemes in classifying gene and protein expression studies. *IEEE Transactions on Computational Biology and Bioinformatics*, 5:368–384, 2008.

[13] J.B. Tenenbaum, V. de Silva, and J.C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290:2319–2323, 2000.

[14] S.T. Roweis and L.K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2323–2326, 2000.

[15] S. Lafon and A. B. Lee. Diffusion maps and coarse-graining: A unified framework for dimensionality reduction, graph partitioning, and data set parameterization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28:1393–1403, 2006.

[16] J. Shi and J. Malik. Normalized Cuts and Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

[17] J. Ham, D.D. Lee, S. Mika, and B. Schölkpof. A kernel view of the dimensionality reduction of manifolds. In *International Conference on Machine Learning*, pages 369–376, 2004.

[18] B. Higgs, J. Weller, and J. Solka. Spectral embedding finds meaningful (relevant) structure in image and microarray data. *BMC Bioinformatics*, 7(1):74, 2006.

[19] P. Tiwari, J. Kurhanewicz, M. Rosen, and A. Madabhushi. Semi supervised multi kernel (sesmik) graph embedding: Identifying aggressive prostate cancer via magnetic resonance imaging and spectroscopy. In *Medical Image Computing and Computer Assisted Intervention*, pages 666–673, 2010.

[20] H. ElGhawalby and E.R. Hancock. Graph embedding using an edge-based wave kernel. In *International Conference on Structral, Syntactic, and Statistical Pattern Recognition*, pages 60–69, 2010.

[21] A. Robles-Kelly and E.R. Hancock. A riemannian approach to graph embedding. *Pattern Recognition*, 40:1042–1056, 2007.

[22] J. Naik, S. Doyle, A. Basavanhally, S. Ganesan, M. Feldman, J. Tomaszeweski, and A. Madabhushi. A boosted distance metric: Application to content based image retrieval and classification of digitized histopathology. In *SPIE Medical Imaging*, 2009.

[23] Y. Freund and R.E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *JCSS: Journal of Computer and System Sciences*, 55:119–139, 1997.

[24] A. Madabhushi, S. Agner, A. Basavanhally, S. Doyle, and G. Lee. Computer-aided prognosis: Predicting patient and disease outcome via quantitative fusion of multi-scale, multi-modal data. *Computerized Medical Imaging and Graphics*, In Press, 2011.

[25] J.P. Monaco, J.E. Tomaszewski, M.D. Feldman, I. Hagemann, M. Moradi, P. Mousavi, A. Boag, C. Davidson, P. Abolmaesumi, and A. Madabhushi. High-throughput detection of prostate cancer in histological sections using probabilistic pairwise markov models. *Medical Image Analysis*, 14(4):617 – 629, 2010.

[26] A. Madabhushi, S. Doyle, G. Lee, A. Basavanhally, J. Monaco, S. Masters, J. Tomaszeweski, and M. Feldman. Review: Integrated diagnostics: A conceptual framework with examples. *Clinical Chemistry and Laboratory Medicine*, 48(7):989–998, 2010.

[27] H. J. G. Bloom and W. W. Richardson. Histological grading and prognosis in breast cancer. *British Journal of Cancer*, 11(3):359–377, September 1957.

[28] C. Kamate, S. Baloul, S. Grootenboer, E. Pessis, A. Chevrot, M. Tulliez, C. Marchiol, M. Viguier, and D. Fradelizi. Inflammation and cancer, the mastocytoma p815 tumor model revisited: triggering of macrophage activation in vivo with pro-tumorigenic consequences. *Int J Cancer*, 100(5):571–579, Aug 2002.

[29] K. Tsuta, G. Ishii, E. Kim, S. Shiono, Y. Nishiwaki, Y. Endoh, T. Kodama, K. Nagai, and A. Ochiai. Primary lung adenocarcinoma with massive lymphocytic infiltration. *American Journal of Clinical Pathology*, 123:547–552, 2005.

[30] G. Alexe, G.S. Dalgin, D. Scanfeld, P. Tamayo, J.P. Mesirov, C. DeLisi, L. Harris, N. Barnard, M. Martel, A.J. Levine, S. Ganesan, and G. Bhanot. High expression of lymphocyte-associated genes in node-negative her2+ breast cancers correlates with lower recurrence rates. *Cancer Res*, 67(22):10669–10676, Nov 2007.

[31] L. Zhang, J.R. Conejo-Garcia, D. Katsaros, P.A. Gimotty, M. Massobrio, G. Regnani, A. Makrigiannakis, H. Gray, K. Schlienger, M.N. Liebman, S.C. Rubin, and G. Coukos. Intratumoral t cells, recurrence, and survival in epithelial ovarian cancer. *N Engl J Med*, 348(3):203–13, 2003.

[32] A. Madabhushi. Digital pathology image analysis: Opportunities and challenges. *Imaging in Medicine*, 1:7–10, 2009.

[33] S Doyle, M Feldman, J Tomaszewski, and A Madabhushi. A boosted bayesian multiresolution classifier for prostate cancer detection from digitized needle biopsies. *IEEE Transactions on Biomedical Engineering,*, (In Press)(99):1, 2010.

[34] B. S. Manjunath and W. Y. Ma. Texture features for browsing and retrieval of image data. *IEEE Transactions Pattern Analysis Machine Intelligence*, 18(8):837–842, August 1996.

[35] A.K. Jain and F. Farrokhnia. Unsupervised texture segmentation using gabor filters. In *IEEE International Conference on Systems, Man and Cybernetics*, 1990.

[36] R.M. Haralick, K. Shanmugam, and I. Dinstein. Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics*, 3(6):610–621, November 1973.

[37] A.N. Basavanhally, S. Ganesan, S. Agner, J.P. Monaco, M.D. Feldman, J.E. Tomaszewski, G. Bhanot, and A. Madabhushi. Computerized image-based detection and grading of lymphocytic infiltration in her2+ breast cancer histopathology. *IEEE Transactions on Biomedical Engineering*, 57(3):642 –653, 2010.

[38] S.A. Hojjatoleslami and J. Kittler. Region growing: A new approach. *IEEE Transactions on Image Processing*, 7:1079–1084, 1995.

[39] J. Besag. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society*, B-48:259–302, 1986.

[40] L. Breiman. Random forests. In *Machine Learning*, pages 5–32, 2001.

[41] R. Sparks and A. Madabhushi. Out-of-sample extrapolation using semi-supervised manifold learning (ose-ssl): Content-based image retrieval for prostate histology grading. In *International Symposium on Biomedical Imaging*, 2011.

Curriculum Vita

Akshay Sridhar

Education

2010 - 2012 M.S. in Biomedical Engineering, Rutgers University

2005 - 2009 B.S. in Biomedical Engineering (Magna Cum Laude), Rutgers University

Work Experience

2009 - 2010 Research Assistant, Department of Biomedical Engineering, Rutgers University

2006 - 2009 Assistant Supervisor, Rutgers Computing Services Help Desk, Rutgers University

Publications

- **Sridhar, A.**, Doyle, S., Madabhushi, A., Content-Based Image Retrieval of Digitized Histopathology in Boosted Spectrally Embedded Spaces, IEEE Transactions on Information Technology in Biomedicine, 2011. (Submitted)
- Tiwari, P., Kurhanewicz, J., Viswanath, S., **Sridhar, A.**, Madabhushi, A., Multimodal Wavelet Embedding Representation for Data Combination (MaWERiC): Integrating Magnetic Resonance Imaging and Spectroscopy for Prostate Cancer Detection, NMR in Biomedicine, 2011 (in press). doi: 10.1002/nbm.1777.