

BOOSTED SPECTRAL EMBEDDING (BOSE): APPLICATIONS TO CONTENT-BASED IMAGE RETRIEVAL OF HISTOPATHOLOGY

Akshay Sridhar, Scott Doyle, Anant Madabhushi

Rutgers University, The State University of New Jersey
Department of Biomedical Engineering
599 Taylor Road, Piscataway, NJ, 08854

ABSTRACT

In machine learning, non-linear dimensionality reduction (NLDR) is commonly used to embed high-dimensional data into a low-dimensional space while preserving local object adjacencies. However, the majority of NLDR methods define object adjacencies using distance metrics that do not account for the quality of the features in the high-dimensional space. In this paper we present Boosted Spectral Embedding (BoSE), a variant of the traditional Spectral Embedding (SE) that utilizes a Boosted Distance Metric (BDM) to improve the low-dimensional representation of the data. Under the naive assumption that all features are equally important, SE uses the Euclidean distance metric to define object-distance relationships. However, the BDM selectively weights features via the AdaBoost algorithm such that the low-dimensional representation contains only the most discriminating features. In this work BoSE is evaluated against SE in the context of digitized histopathology images using (a) content-based image retrieval and (b) classification via Random Forest of the low-dimensional representation. Using images from a cohort of 58 prostate cancer patient studies, BoSE and SE separated benign and malignant samples with areas under the precision-recall curve (AUPRCs) of 0.95 and 0.67 and classification accuracies using a Random Forest (RF) classifier were 0.93 and 0.79, respectively. For a cohort of 55 breast cancer studies, BoSE and SE performed comparably in terms of both RF accuracy and AUPRC. In addition, a qualitative visualization of the low-dimensional data representations suggests that BoSE exhibits improved class separability over SE.

Index Terms— content-based image retrieval, spectral embedding, boosting, histopathology, BoSE, prostate cancer, breast cancer

1. INTRODUCTION

Non-linear dimensionality reduction (NLDR) methods such as Isomaps, Locally Linear Embedding, and Spectral Embedding (SE) are able to reduce the dimensionality of large datasets [1]. Although NLDR methods have different motivations and derivations, each method attempts to map high-dimensional data into a lower dimensional space while preserving the local neighborhood structure. These techniques are extremely useful in their ability to reduce the complexity of analysis while mitigating the curse of dimensionality. For example, SE defines object adjacency by using

This work was made possible by the Wallace H. Coulter Foundation, New Jersey Commission on Cancer Research, National Cancer Institute (R01CA136535-01, R01CA140772-01, R21CA127186-01, and R03CA143991-01), and The Cancer Institute of New Jersey. We also wish to thank Dr. John Tomaszewski, Dr. Michael Feldman, Dr. Carolyn Mies, and Dr. Shridar Ganesan for providing and annotating the digitized histopathology data.

a Gaussian kernel in conjunction with the Euclidean distance metric to yield a confusion matrix for all objects. A key shortcoming of the Euclidean distance metric, however, is that it implicitly assumes all features (dimensions) are equally relevant and weights all elements of the confusion matrix accordingly.

In this paper we introduce a novel variant of SE called Boosted Spectral Embedding (BoSE) that utilizes a Boosted distance metric (BDM) in lieu of the traditional Euclidean distance. The BDM is a modified version of the metric proposed in [2]. The BDM employs AdaBoost [3], a popular classifier ensemble scheme, to select features based on their ability to discriminate between two classes and assign feature weights accordingly. Subsequently, the Gaussian kernel used in SE is modified to utilize the BDM in place of Euclidean distance. The difference between SE and BoSE is that BDM actively places importance on discriminatory features while downplaying the role of weaker features.

There has been considerable previous work in the development of SE variants. Xiong et al. [4] developed a boosted graph embedding scheme for the application of facial expression recognition where they used a kernel that incorporated the weights of the samples found via AdaBoost.M2. ElGhawalby et al. [5] 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 pseudo-Riemannian manifold in a pseudo-Euclidean space and used the resulting point-set to compute the graph characteristics. Robles-Kelly et al. [6] used the Kruskal coordinates to compute the edge-weights for the weight matrix and used it to embed the nodes of the graph onto a Riemannian manifold. Tiwari et al. [7] developed a method that used multiple kernels to develop the weight matrix. However, to the best of our knowledge, there has not been any work done that has leveraged a learned distance metric in conjunction with SE.

In order to quantitatively evaluate BoSE, we compare it against SE for content-based image retrieval (CBIR) in the context of digitized histopathology. The application of BoSE to CBIR is an important task since Doyle et al. [8] 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. In addition, Reddy et al. [9] and Yang et al. [10] have shown that the use of a learned distance metric for the purpose of medical image retrieval yields a larger number of relevant images returned.

In this work, CBIR systems are constructed for distinguishing histopathology images containing (a) low and high grade breast cancers and (b) benign and malignant prostate cancers. All the retrieval is performed in the reduced dimensional space generated by BoSE and SE, respectively. For a set of N images, features were extracted for each image to form a high dimensional data set. The data was then embedded into a three dimensional space using BoSE and SE

to create the manifolds $\mathcal{M}^{\text{BoSE}}$ and \mathcal{M}^{SE} . Each of the N images is successively treated as the query image and the closest images on $\mathcal{M}^{\text{BoSE}}$ and \mathcal{M}^{SE} are then retrieved. Precision recall (PR) curves are then generated for each of BoSE and SE and the corresponding areas under the PR curves are compared, larger area signifying improved performance. A second performance measure for evaluating BoSE and SE is via classifier accuracy. In this work a Random Forest (RF) classifier (obtained by Bagging decision trees) was trained on both $\mathcal{M}^{\text{BoSE}}$ and \mathcal{M}^{SE} . The accuracy of the RF classifier should reflect the quality of $\mathcal{M}^{\text{BoSE}}$ and \mathcal{M}^{SE} and hence the class discriminability offered by the BoSE and SE schemes, respectively.

The primary contributions of this work are:

- The construction of a novel kernel representation for SE that uses Boosting [3] to identify which individual features should be used for embedding high-dimensional data into a low-dimensional space.
- A reduced dimensional manifold representation that yields greater class discriminability compared to the traditional SE scheme.
- In conjunction with BoSE, the construction of a CBIR system that is employed for digital histopathology.

The rest of the paper is organized as follows. The formulations for SE and BDM are presented in Sections 2 and 3, respectively. The theory and formulation for BoSE is presented in Section 4 and the application of BoSE to CBIR is presented in Section 5. The experimental results and discussion are presented in Section 6. Lastly, concluding remarks are presented in Section 7.

2. SPECTRAL EMBEDDING FORMULATION

Let $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\} \in \mathbb{R}^{N \times D}$ where \mathbf{X} contains N objects with D dimensions. The goal of SE is to project the feature vectors from a D dimensional space on to a k dimensional space where $k \ll D$. Let $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_N\} \in \mathbb{R}^{N \times k}$ where \mathbf{Y} is the low dimensional projection of \mathbf{X} . The eigenvector \mathbf{y}_i associated with object $i \in \{1, 2, \dots, N\}$ is defined as

$$\mathbf{y} = \underset{\mathbf{y}}{\operatorname{argmin}} \left(\sum_{i,j=1}^N \|y_i - y_j\|^2 w_{ij} \right), \quad (1)$$

where $w_{ij} = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/\sigma)$ is the (i, j) element of the weight matrix $\mathbf{W} \in \mathbb{R}^{N \times N}$ that assigns edge weights to characterize pairwise similarities between objects \mathbf{x}_i and \mathbf{x}_j for $i, j \in \{1, 2, \dots, N\}$. The constant σ is used as a scaling parameter. The minimization of Equation (1) reduces to an eigenvalue decomposition problem, $(L - \mathbf{W})\mathbf{y} = \lambda L\mathbf{y}$, where L is the diagonal matrix, $L_{ii} = \sum_i w_{ij}$.

3. BOOSTED DISTANCE METRIC

Let $\Phi_d(\mathbf{x}_i)$ for $d \in \{1, 2, \dots, D\}$ represent the value of feature d from \mathbf{x}_i . Let $\mathcal{L}(\mathbf{x}_i) \in \{\omega_1, \omega_2\}$ represent the ground truth label denoting positive or negative class, respectively. The BDM is constructed in three main steps. First, for each d , we construct a weak Bayesian classifier $h_d(\Phi(\mathbf{x}_i)) \in \{\omega_1, \omega_2\}$, $d \in \{1, 2, \dots, D\}$ to perform class prediction for \mathbf{x}_i . Next, the T most accurate classifiers $h_t, t \in \{1, 2, \dots, T\}$ are identified and weights associated with each h_t are learned via AdaBoost [3]. The BDM is then defined using the features Φ_t and associated weights α_t .

3.1. Construction of Weak Classifiers

Step 1: Extract quantitative features [2] from all $\mathbf{x}_i \in \mathbb{R}^{1 \times D}$.

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

Step 3: Let $\Phi_d(\mathbf{X})$ denote the vector containing the values of feature $d \in \{1, 2, \dots, D\}$. $\Phi_d^+(\mathbf{X})$ and $\Phi_d^-(\mathbf{X})$ are the feature values from the positive and negative class respectively. Obtain probability distribution functions for both \mathcal{B}_d^+ and \mathcal{B}_d^- by solving,

$$p(\Phi_d(\mathbf{X}), \omega_a) = \Phi_d^{\tau-1}(\mathbf{X}) \frac{\exp\left(\frac{-\Phi_d(\mathbf{X})}{\eta}\right)}{\eta^\tau \Gamma(\tau)} \quad (2)$$

where Γ is the gamma function that is parameterized by a scale parameter τ and a shape parameter η with $\tau, \eta > 0$.

Step 4: Obtain the *a posteriori* probability $P(\omega_1|\Phi_d(\mathbf{X}))$ which computes the likelihood that an object with feature value Φ_d belongs to the positive class ω_1 by solving,

$$P(\omega_1|\Phi_d(\mathbf{X})) = \frac{P(\omega_1)p(\Phi_d(\mathbf{X})|\omega_1)}{\sum_a P(\omega_a)p(\Phi_d(\mathbf{X})|\omega_a)} \quad (3)$$

3.2. Learning Feature Weights via AdaBoost

We use the AdaBoost algorithm [3] to select the weak classifiers that are best able to distinguish between the positive and negative classes. AdaBoost works in an iterative fashion by choosing the best-performing weak classifiers and assigning weights based on the classification accuracy of that feature. All samples are initialized with equal weights. At each iteration, weights increase for samples that are difficult to classify. This ensures that subsequent weak classifiers focus on these samples. The output of the algorithm is a set of weak classifiers h_g and their associated weights $\alpha_g, g \in \{1, 2, \dots, G\}$ where $1 \leq G \leq D$ and $\alpha_g > 0$.

3.3. 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,

$$\mathbb{D}_{\text{BDM}}(\mathbf{x}_i, \mathbf{x}_j) = \left[\sum_{g=1}^G \alpha_g (\Phi_g(\mathbf{x}_i) - \Phi_g(\mathbf{x}_j))^2 \right]^{\frac{1}{2}} \quad (4)$$

Proposition: Given that $\mathbb{D}_{\text{Eu}} = \left[\sum_{g=1}^G (\Phi_g(\mathbf{x}_i) - \Phi_g(\mathbf{x}_j))^2 \right]^{\frac{1}{2}}$ is the Euclidean distance metric, \mathbb{D}_{BDM} is also a distance metric.

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

4. BOOSTED SPECTRAL EMBEDDING (BOSE)

The Boosted Spectral Embedding (BoSE) is a variant of SE where the computation of the confusion matrix \mathbf{W} is defined as,

$$w_{ij} = \exp\left(-\frac{\mathbb{D}_{\text{BDM}}(\mathbf{x}_i, \mathbf{x}_j)}{\sigma}\right). \quad (5)$$

Data	Classes (ω_1/ω_2)	Features	Class Distribution (ω_1/ω_2)
Breast	High Grade/Low Grade	Haralick	36/19
Prostate	Cancer/Benign	Gabor, Haralick, first-order statistics	29/29

Table 1. List of the breast cancer and prostate cancer datasets used in this study.

Since \mathbb{D}_{BDM} is a metric, \mathbf{W} is positive, semi-definite, and symmetric. The idea behind using the BDM to define \mathbf{W} is to place importance on features that are good at discriminating between objects of different classes. This results in a final embedding that contains greater separation between objects from different classes as compared to the Euclidean distance.

5. EVALUATION OF BOSE USING CBIR AND CLASSIFIER ACCURACY

5.1. BoSE for Content-Based Image Retrieval

Algorithm: *BoSE*

Input: Training samples \mathcal{S}^t , Testing samples \mathcal{S}^{te} , $\mathcal{L}\{\mathcal{S}^{tr}\}$, $\mathcal{L}\{\mathcal{S}^{te}\}$, iterations G

Output: Lower dimensional embedding \mathbf{Y}

begin

- 1: Build weak classifiers $h_d : d \in \{1, 2, \dots, D\}$ via a Bayesian Classifier;
- 2: Select optimal weak classifiers h_g and weights α_g for $g \in \{1, 2, \dots, G\}$ via AdaBoost;
- 3: Obtain BDM by applying Equation 4;
- 4: Obtain W by Equation 5;
- 5: Find $\mathbf{Y} \in \mathbb{R}^{N \times k}$ by Equation 1;
- 6: Output \mathbf{Y} .

end

5.2. Precision Recall Curves

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. Euclidean distance within an embedding defines object adjacency. Recall is defined as $r(\beta) = \frac{\xi(\beta)}{\xi(N)}$. Precision recall curves were generated by plotting $\rho(\beta)$ versus $r(\beta)$ for $\beta \in \{1, 2, \dots, N\}$. Area under the precision recall curve (AUPRC) was measured and used to evaluate the CBIR system.

5.3. Classifier Accuracy

The discriminability of the reduced dimensional manifolds $\mathcal{M}^{\text{BoSE}}$ and \mathcal{M}^{SE} (Figure 1) is evaluated directly via a Random Forest (RF) classifier. A RF classifier is an ensemble of decision trees (i.e. weak learners) combined via bootstrap aggregation. By averaging decisions across the weak classifiers, a strong classifier that reduces over-all bias and variance is created.

A randomized, 3-fold cross-validation scheme was used to determine training and testing inputs for the RF classifier. First, the entire dataset is randomly divided into three equally-sized subsets. Two of the subsets are used for training the RF classifier, which is then evaluated on the remaining subset. The subsets are subsequently rotated until each subset is used for evaluation exactly once. The entire

cross-validation scheme is repeated over 50 iterations, over which mean and standard deviation classification accuracy are reported.

6. EXPERIMENTAL DESIGN AND RESULTS

6.1. Dataset Description

For this paper we considered two datasets comprising (1) 55 breast and (2) 58 prostate biopsy tissue samples, respectively (Table 1). Slides from both 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. Ground truth annotations were delineated by an expert pathologist for all data and Bloom-Richardson (BR) grades accompanied all breast cancer data.

6.2. Experiment 1: Distinguishing Malignant from Benign Prostate Histopathology

Due to the large volume of prostate cancer biopsy samples, manual inspection is a time-consuming and labor-intensive process. Hence the automated detection of malignant from benign tissue on prostate biopsy samples is an important task. The objective of this experiment was to distinguish between malignant and benign prostate tissue patches (Table 1) from biopsy samples obtained from 55 patients. From each image, 14 texture features (Gabor, Haralick, first-order statistics) were extracted from each image [8] and Adaboost identified and weighted the top 4 performing features. AUPRC for CBIR with embeddings generated by BoSE (CBIR-BoSE) was 0.95 ± 0.01 (Table 2). AUPRC for CBIR with embeddings generated by SE (CBIR-SE) was 0.67 ± 0.13 . Figure 2(a) displays the precision recall curves for CBIR-BoSE and CBIR-SE demonstrating CBIR-BoSE performs considerably better than CBIR-SE. Figures 1(a) and 1(b) displays the embeddings for BoSE and SE, respectively, showing a much greater separation between the malignant and benign classes in the low-dimensional embedding plots created by BoSE compared to SE.

Data	Method	AUPRC	RF Accuracy
Prostate Cancer Detection	BoSE	0.95 ± 0.01	0.93 ± 0.01
	SE	0.67 ± 0.13	0.79 ± 0.01
Breast Cancer Grading	BoSE	0.88 ± 0.02	0.98 ± 0.01
	SE	0.88 ± 0.02	0.98 ± 0.02

Table 2. Quantitative results for Boosted Spectral Embedding (BoSE) and traditional Spectral Embedding (SE) are presented for both the breast and prostate cancer datasets. For all experiments, embedding performance was measured via (1) area under the precision recall curve (AUPRC) and (2) cross-validation accuracy of a Random Forest (RF) classifier shown.

6.3. Experiment 2: Distinguishing High from Low Grade Breast Histopathology

Manual grading of breast cancer suffers from high inter- and intra-pathologist variability. Hence the automated grading of breast cancer

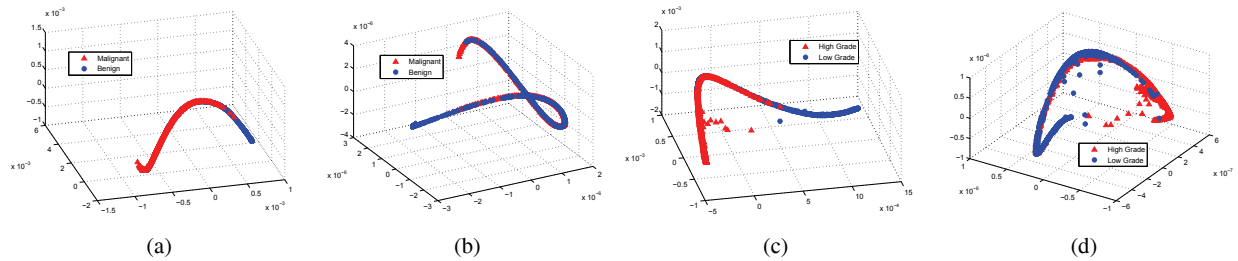


Fig. 1. Low-dimensional embedding plots are shown illustrating (a), (b) prostate and (c), (d) breast cancer histopathology images using (a), (c) BoSE and (b), (d) SE methods. It is clear from visual inspection that the embeddings produced by BoSE yield better class separation as compared to SE.

histopathology could potentially enable the creation of a quantitative, reproducible measure of tumor malignancy. 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 1). From each image, 50 Haralick features [8] are extracted to describe textural patterns. Adaboost identified and weighted 12 optimal features. CBIR-BoSE had an AUPRC of 0.88 ± 0.02 compared to CBIR-SE yielded an AUPRC of 0.88 ± 0.02 (Table 2). The intersection of the PR curves shown in Figure 2(b) indicates that they have the same performance at that point. Overall, however, CBIR-BoSE outperforms CBIR-SE. The embedding produced by BoSE (Figure 1(c)) shows improved separation between the low and high grade classes over the embedding produced by SE (Figure 1(d)). While no differences were observed between the AUPRC and the classifier accuracy values for the SE and BoSE scheme, in this application, the lower dimensional embedding representations (Figures 1(c),(d)) appear to suggest better class separability between the two classes in the BoSE space compared to SE.

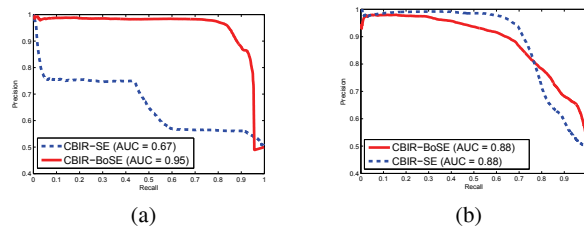


Fig. 2. Precision-recall curves for (a) prostate and (b) breast histopathology datasets. Note that while similar AUPRC values were obtained for the breast cancer dataset for the BoSE and SE schemes, significant improvement was observed for the BoSE scheme for the prostate dataset.

7. CONCLUDING REMARKS

In this paper, we presented Boosted Spectral Embedding (BoSE), which modifies the traditional Spectral Embedding (SE) to use the improved Boosted Distance Metric (BDM). The BDM preferentially weights features that discriminate between objects of different classes allowing for a confusion matrix which better describes object similarity. By altering the confusion matrix BoSE yields an embedding with improved separation between objects of differing classes. We compared BoSE to the Spectral Embedding (SE)

scheme, which uses the Euclidean distance metric to define object similarity. Quantitative results showed that BoSE had a higher AUPRC and classification accuracy compared to SE for retrieving benign or malignant prostate histopathology images. For breast cancer grading, BoSE achieved identical results (via AUPRC and classification accuracy measures) to SE. This may have been due to the fact that the features used for BoSE and SE for this dataset were highly discriminating to begin with. The reduced dimensional embedding for this dataset, however, appeared to show greater class discriminability for BoSE compared to SE.

8. REFERENCES

- [1] G. Lee et al., "Investigating the efficacy of nonlinear dimensionality reduction schemes in classifying gene and protein expression studies," *IEEE Trans. CBIB*, vol. 5, pp. 368–384, 2008.
- [2] J. Naik et al., "A boosted distance metric: application to content based image retrieval and classification of digitized histopathology," in *SPIE Medical Imaging*, 2009.
- [3] 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*, vol. 55, pp. 119–139, 1997.
- [4] L. Xiong et al., "Boosting graph embedding with application to facial expression recognition," in *IAS*, 2009, pp. 707–710.
- [5] H. ElGhawalby et al., "Graph embedding using an edge-based wave kernel.," in *SSPR/SPR*, 2010, pp. 60–69.
- [6] A. Robles-Kelly et al., "A riemannian approach to graph embedding," *Pattern Recognition*, vol. 40, pp. 1042–1056, 2007.
- [7] P. Tiwari et al., "Semi supervised multi kernel (sesmik) graph embedding: Identifying aggressive prostate cancer via magnetic resonance imaging and spectroscopy," in *MICCAI*, 2010, pp. 666–673.
- [8] S. Doyle et al., "Using manifold learning for content-based image retrieval of prostate histopathology," in *MICCAI*, 2007, pp. 53–62.
- [9] C.K. Reddy et al., "Retrieval and ranking of biomedical images using boosted haar features," in *BIBE*, 2008, pp. 1–6.
- [10] L. Yang et al., "A boosting framework for visuality-preserving distance metric learning and its application to medical image retrieval," *IEEE PAMI*, vol. 32, pp. 30–44, 2010.