Automated Discrimination of Shapes in High Dimensions

Linh Lieu and Naoki Saito

Dept. of Mathematics, Univ. of California-Davis, One Shields Ave., Davis, CA, USA 95616

ABSTRACT

We present a new method for discrimination of data classes or data sets in a high-dimensional space. Our approach combines two important relatively new concepts in high-dimensional data analysis, i.e., Diffusion Maps and Earth Mover's Distance, in a novel manner so that it is more tolerant to noise and honors the characteristic geometry of the data. We also illustrate that this method can be used for a variety of applications in high dimensional data analysis and pattern classification, such as quantifying shape deformations and discrimination of acoustic waveforms.

Keywords: Shape discrimination, high-dimensional pattern classification, high-dimensional histograms, dimensionality reduction, Diffusion Maps, Earth Mover's Distance

1. INTRODUCTION

In many applications, we would want to automatically discriminate one class of signals or one set of data from another. For example, sonar signal processing applications require the recognition of different classes of acoustic waveforms reflected from different types of underwater objects. In medical image analysis, it is desirable to automatically quantify shape deformations of a certain brain structure in order to track the development of certain illnesses. Tracking the size and shape changes of the corpus callosum at different stages of cognitive development in children is such an example.

The main challenge in these discrimination problems is the high-dimensional nature of the data, the *curse of dimensionality.* The data points are often recorded in a high-dimensional ambient space, but the majority of the points lie in a low-dimensional subspace of this ambient space. Sometimes the data may be in a low-dimensional space but consist of a large number of points with high redundancy. In the first case, the multidimensional scalings of a few data points completely obscure the intrinsic geometric structure of the data. In the second case, it is quite inefficient to process all data points without considering the high redundancy in the data. A solution to the high dimensionality problem is to find a low-dimensional representation for the data, i.e., to perform dimensionality reduction, and then subsample the data in such a way that intrinsic geometric structures are preserved.

In recent years, R. R. Coifman and S. Lafon introduced the method of *Diffusion Geometry* for dimensionality reduction.[1,2,3] This method extends an earlier idea of using Laplacian eigenmaps for data representation by M. Belkin and P. Niyogi[4] and proves to be quite successful in achieving dimension reduction while preserving the intrinsic geometries of the data.

In this paper, we propose a new method for discrimination of data classes or sets in high dimensions. The first step in our discrimination approach is to utilize the techniques of Diffusion Geometry to project or embed the data into a low-dimensional space. In this space the intrinsic geometries of the data are preserved and highlighted. We then extract these highlighted geometric structures from the data to form discriminant features for each data set.

Once discriminant features are found, the next step is to define a discriminant measure on the sets of discriminant features. We want a measure that takes in any two sets of discriminant features and outputs a real number on the scale from 0 to 1, where 0 means that the two data classes are the same and 1 means that they are extremely different. For this purpose, we utilize the *Earth Mover's Distance* in the framework of Y. Rubner

Further author information: (Send correspondence to L.L.)
L.L.: E-mail: llieu@math.ucdavis.edu, URL: http://www.math.ucdavis.edu/~llieu
N.S.: E-mail: saito@math.ucdavis.edu, URL: http://www.math.ucdavis.edu/~saito

and C. Tomasi[5,6] as our discriminant measure. Under certain conditions, the Earth Mover's Distance is a metric between two distributions. In a probabilistic interpretation, the Earth Mover's Distance is equivalent to the Mallows distance on probability distributions.[7] For our case, each set of discriminant features (hereafter to be called a *signature*) can be viewed as a discrete probability distribution. So the Earth Mover's Distance between two signatures is essentially a distance between two discrete probability distributions.

The organization of this paper is as follows: we first review the important concepts and techniques of Diffusion Geometry and Earth Mover's Distance. Then we describe our proposed method for discrimination in Sec. 4, following are illustrations of two applications in Sec. 5.1 and Sec. 5.2. One application is discrimination of different classes of sonar signals, and the other is quantification of the differences between three different corpus callosum shapes segmented from the sagittal view of three MRI scans.

2. REVIEW OF DIFFUSION MAPS

Reduction in the dimensionality of the data can be achieved by utilizing *diffusion maps* to embed the data into a lower-dimensional space called *diffusion space*.[1,2] These diffusion maps are constructed from the eigenfunctions of a *diffusion operator* whose kernel is called a *diffusion kernel*. One way to construct a diffusion kernel from the data is as follows:

Suppose the data set X belongs to a space having a natural dissimilarity measure d that gives a sense of affinity between any two points in X. This is a reasonable assumption to make in practice. For example, if X is a database of images, then d may be the L^2 norm between two images. Or, if X belongs to a submanifold in $\mathbb{R}^n$, then d may be the usual Euclidean distance.

For $\epsilon > 0$, let $w_\epsilon(\boldsymbol{x}, \boldsymbol{y}) := \mathrm{e}^{-(d(\boldsymbol{x},\boldsymbol{y})/\epsilon)^2}$. It will soon become clear that this w_ϵ gives the notion of local geometry to X. That is, it defines the notion of a local neighborhood at each point $\boldsymbol{x} \in X$ via the affinity between $\boldsymbol{x}$ and other points, and the value of the parameter ϵ specifies the size of this neighborhood. By renormalizing w_ϵ to be row stochastic (to have sum 1 along the $\boldsymbol{y}$ direction), we obtain the diffusion kernel

$$k(\boldsymbol{x}, \boldsymbol{y}) := \frac{w_\epsilon(\boldsymbol{x}, \boldsymbol{y})}{p_\epsilon(\boldsymbol{x})},$$

where $p_\epsilon(\boldsymbol{x}) := \sum_{\boldsymbol{y} \in X} w_\epsilon(\boldsymbol{x}, \boldsymbol{y})$. The corresponding diffusion operator is

$$Af(\boldsymbol{x}) := \sum_{\boldsymbol{y} \in X} k(\boldsymbol{x}, \boldsymbol{y}) f(\boldsymbol{y}).$$

The kernel k is non-negative and row-stochastic, therefore it can be viewed as a transition matrix of a Markov process on X. The operator A is an averaging operator, since it is positivity-preserving (if $f \geq 0$ then $Af \geq 0$) and preserves constant functions. In other words, the action of the operator A can be interpreted as 'diffusion' of information, and the Markov chain specifies fast and slow directions of propagation and accumulation. Observe that the probability of transition between two points is high if they have strong affinity, thus information flows and accumulates in regions of high affinity.

One main idea in the framework of diffusion geometry is to take larger powers of the operator A. This can be interpreted as running the Markov process forward in time, or equivalently, letting information diffuse for longer time. Since information flows with respect to the affinity between points, as time passes, information accumulates in regions of high affinity. This is essentially how the local geometry in the data is captured in the diffusion geometry framework. Now, for $t > 0$, let $k^{(t)}$ denotes the kernel of A^t. Note that $k^{(t)}(\boldsymbol{x}, \boldsymbol{y})$ is the probability of transition from $\boldsymbol{x}$ to $\boldsymbol{y}$ in t steps. The *diffusion distance* between two points $\boldsymbol{x}$ and $\boldsymbol{y}$ is defined as

$$D_t(\boldsymbol{x}, \boldsymbol{y})^2 := \|k^{(t)}(\boldsymbol{x}, \cdot) - k^{(t)}(\boldsymbol{y}, \cdot)\|^2_{L^2(X, d\mu/v^2)} = \sum_{\boldsymbol{z} \in X} \frac{(k^{(t)}(\boldsymbol{x}, \boldsymbol{z}) - k^{(t)}(\boldsymbol{y}, \boldsymbol{z}))^2}{(v(\boldsymbol{z}))^2},$$

where the "weight" $(v(\boldsymbol{z}))^2 := \frac{p_\epsilon(\boldsymbol{z})}{\sum_{\boldsymbol{z}' \in X} p_\epsilon(\boldsymbol{z}')}$ penalizes discrepancies on regions of low affinity more than those of high affinity.

On an intuitive level, the diffusion distance $D_t(\boldsymbol{x}, \boldsymbol{y})$ measures the affinity between two points based on the transition probabilities in the Markov chain. $D_t(\boldsymbol{x}, \boldsymbol{y})$ is small if there is a large probability of transition from $\boldsymbol{x}$ to $\boldsymbol{y}$ and vice versa. Notice also that in its definition, $D_t(\boldsymbol{x}, \boldsymbol{y})$ takes into account all incidences relating $\boldsymbol{x}$ and $\boldsymbol{y}$. Hence, it is robust to noise perturbations. Consequently, it is a great measure to use for extracting local geometric features from high-dimensional data containing low-dimensional geometric structures.

An approximation to $D_t(\boldsymbol{x}, \boldsymbol{y})$ may be more practical in practice. This is done by considering the spectral decomposition of the kernel k. For this purpose, conjugate k by v to obtain a symmetric kernel:

$$\tilde{k}(\boldsymbol{x}, \boldsymbol{y}) := v(\boldsymbol{x})k(\boldsymbol{x}, \boldsymbol{y})\frac{1}{v(\boldsymbol{y})} = \frac{w_\epsilon(\boldsymbol{x}, \boldsymbol{y})}{\sqrt{p_\epsilon(\boldsymbol{x})}\sqrt{p_\epsilon(\boldsymbol{y})}}.$$

This new kernel shares the same spectrum as k, and its eigenfunctions are obtained via conjugation by v. The operator with kernel $\tilde{k}$:

$$\tilde{A}f(\boldsymbol{x}) := \sum_{\boldsymbol{y}\in X} \tilde{k}(\boldsymbol{x}, \boldsymbol{y})f(\boldsymbol{y})$$

is symmetric and positive semi-definite. Moreover, it is compact with $||\tilde{A}|| = 1$ achieved by eigenfunction v, i.e., $\tilde{A}v = v$. Therefore, $\tilde{A}$ has a discrete, non-increasing, non-negative spectrum: $\lambda_0 = 1 > \lambda_1 \geq \lambda_2 \geq \cdots \geq 0$, and the kernel $\tilde{k}$ has spectral decomposition

$$\tilde{k}(\boldsymbol{x}, \boldsymbol{y}) = \sum_{j\geq 0} \lambda_j \phi_j(\boldsymbol{x})\phi_j(\boldsymbol{y})$$

where $\{\phi_j\}$ is an orthonormal set of eigenfunctions of $\tilde{A}$ forming a basis of $L^2(X)$.

Notice that $v(\boldsymbol{x}) = \phi_0(\boldsymbol{x})$ which is the stationary distribution of the Markov chain, and $\psi_j(\boldsymbol{x}) := \frac{\phi_j(\boldsymbol{x})}{v(\boldsymbol{x})}$ and $\varphi_j(\boldsymbol{y}) := v(y)\phi_j(\boldsymbol{y})$ are left and right (respectively) eigenfunctions of the operator A, and

$$k(\boldsymbol{x}, \boldsymbol{y}) = \sum_{j\geq 0} \lambda_j \psi_j(\boldsymbol{x})\varphi_j(\boldsymbol{y}).$$

Therefore, the kernel of the operator A^t is

$$k^{(t)}(\boldsymbol{x}, \boldsymbol{y}) = \sum_{j\geq 0} \lambda_j^t \psi_j(\boldsymbol{x})\varphi_j(\boldsymbol{y}).$$

Consequently, since $\{\phi_j\}$ are orthonormal, the diffusion distance can be written as

$$D_t(\boldsymbol{x}, \boldsymbol{y})^2 = \sum_{j\geq 1} \lambda_j^{2t}(\psi_j(\boldsymbol{x}) - \psi_j(\boldsymbol{y}))^2.$$

The non-increasing property of the spectrum implies that for any $\delta > 0$, by taking $s(\delta, t) := \max\{j \in \mathbb{N} : |\lambda_j|^t > \delta|\lambda_1|^t\}$, the diffusion distance can be approximated to a relative accuracy δ by

$$D_t(\boldsymbol{x}, \boldsymbol{y})^2 \approx \sum_{j=1}^{s(\delta,t)} \lambda_j^{2t}(\psi_j(\boldsymbol{x}) - \psi_j(\boldsymbol{y}))^2.$$

From this, the diffusion maps are defined as

$$\Psi_t : \boldsymbol{x} \to \begin{pmatrix} \lambda_1^t \psi_1(\boldsymbol{x}) \\ \lambda_2^t \psi_2(\boldsymbol{x}) \\ \vdots \\ \lambda_{s(\delta,t)}^t \psi_{s(\delta,t)}(\boldsymbol{x}) \end{pmatrix}.$$

These diffusion maps can be viewed as coordinates in a $s(\delta, t)$-dimensional Euclidean space characterized by the parameters ϵ, t, and δ. We shall call this space a *diffusion space*.

Via the diffusion maps we have an embedding of the data into the diffusion space denoted by $\mathbb{R}^{s(\delta,t)}$. Moreover, the usual Euclidean distance in this diffusion space is an approximation to the diffusion distance. The key point here is that the diffusion maps give a low-dimensional representation of the data that highlights the underlying intrinsic geometries in the data.

3. REVIEW OF EARTH MOVER'S DISTANCE

The definition of the Earth Mover's Distance (EMD) is based on the solution to a discrete optimal mass transportation problem. Basically, EMD represents the minimum cost of moving earth (or sand) from some source locations to fill up holes at some sink locations. In other words, given any two distributions, one can be viewed as a distribution of earth and the other a distribution of holes, then EMD between the two distributions is the minimum cost of rearranging the mass in one distribution to obtain the other. In the continuous setting, this problem is known as the Monge-Kantorovich optimal mass transfer and has been well studied over the past 100 years. (For an introductory reading on the problem, see Ref. 8.) The importance here is that EMD can be applied to measure the discrepancy between two multidimensional distributions.

In the discrete setting, the optimal mass transfer problem can be formulated as a linear optimization problem as follows:[5,6] Suppose we have a source mass distribution $P = \{(\boldsymbol{p}_1, w_{\boldsymbol{p}_1}), \cdots, (\boldsymbol{p}_m, w_{\boldsymbol{p}_m})\}$ and a sink distribution $Q = \{(\boldsymbol{q}_1, w_{\boldsymbol{q}_1}), \cdots, (\boldsymbol{q}_n, w_{\boldsymbol{q}_n})\}$ in some high-dimensional space $\mathbb{R}^s$. Notice that in this setting P and Q can be viewed as two signatures (as defined in Ref. 5) containing m and n clusters with representatives $\boldsymbol{p}_i$, $\boldsymbol{q}_j \in \mathbb{R}^s$ and weights $w_{\boldsymbol{p}_i} \geq 0$, $w_{\boldsymbol{q}_j} \geq 0$, respectively. Suppose the cost of moving one unit of mass from $\boldsymbol{p}_i$ to $\boldsymbol{q}_j$ is c_{ij}, and f_{ij} denotes the amount of mass flow from $\boldsymbol{p}_i$ to $\boldsymbol{q}_j$. The optimal mass transfer problem is to find the flow $\boldsymbol{F} = [f_{ij}]$ that transfers the maximum allowable amount of earth to fill up the holes with minimum total transportation cost, i.e.,

$$\min_{\boldsymbol{F}} \mathrm{COST}(P, Q, \boldsymbol{F}) := \sum_{i=1}^{m} \sum_{j=1}^{n} c_{ij} f_{ij},$$

subject to

(i) $f_{ij} \geq 0$, for all i, j;

(ii) $\sum_{j=1}^{n} f_{ij} \leq w_{\boldsymbol{p}_i}$, for all $1 \leq i \leq m$;

(iii) $\sum_{i=1}^{n} f_{ij} \leq w_{\boldsymbol{q}_j}$, for all $1 \leq j \leq n$; and

(iv) $\sum_{i=1}^{n} \sum_{j=1}^{n} f_{ij} = \min\left(\sum_i w_{\boldsymbol{p}_i}, \sum_j w_{\boldsymbol{q}_j}\right)$.

Constraint (i) ensures that one can only move earth from P to Q, not vice versa; (ii) that the amount of earth moved from P is no more than the sum of the weights $w_{\boldsymbol{p}_i}$; (iii) that the amount of earth received at Q is no more than the sum of the weights $w_{\boldsymbol{q}_j}$; and (iv) that the maximum allowable amount of earth is moved.

Once the optimal flow $\boldsymbol{F}^*$ from P to Q is found, EMD is then defined as the total cost normalized by the total flow:

$$\mathrm{EMD}(P, Q) := \frac{\mathrm{COST}(P, Q, \boldsymbol{F}^*)}{\sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij}^*} = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} c_{ij} f_{ij}^*}{\sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij}^*}.$$

Notice that the normalization factor is the total weight of the smaller signature (due to constraint (iv)). This normalization ensures that smaller signatures are not favored in the case when two signatures have different total weights. Furthermore, EMD is symmetric, i.e., $\mathrm{EMD}(P, Q) = \mathrm{EMD}(Q, P)$ for any two distributions P and Q.

4. OUR PROPOSED METHOD

In this section, we describe how diffusion maps and Earth Mover's Distance can be applied to do discrimination tasks. Our approach quantitatively determines the similarity or dissimilarity between classes of data of high dimensional nature. We first describe in details our proposed method, and then give some examples of application.

Given $X_1, \cdots, X_m$ data sets or classes, our approach for discrimination involves the following steps:

1. Signature construction in diffusion space.
   i. Construct diffusion maps Ψ_t on $X = \cup_{j=1}^m X_j$ and embed X into a diffusion space $\mathbb{R}^{s(\delta,t)}$ using Ψ_t.
   ii. Denote $Y_j := \Psi_t(X_j)$, the set of (embedded) points in diffusion space $\mathbb{R}^{s(\delta,t)}$ corresponding to X_j. For each X_j $(j = 1, \cdots, m)$, construct its signature SG_j by applying Lafon-Lee's Coarse-Graining algorithm[3] to Y_j.
2. Computation of discriminant measure.
   i. Compute Earth Mover's Distance between pairs of signatures SG_i and SG_j $(i, j = 1, \cdots, m)$, then define this to be the discriminant measure between the data classes X_i and X_j.

A signature is essentially a high-dimensional version of a histogram. As mentioned in Sec. 3, each signature SG_j consists of k_j clusters of points. Each cluster has a representative $\boldsymbol{sg}_{j,\ell}$ and carries a weight $w_{\boldsymbol{sg}_{j,\ell}} \geq 0$ $(\ell = 1, \cdots, k_j)$. For our case, each $\boldsymbol{sg}_{j,\ell}$ belongs to the diffusion space $\mathbb{R}^{s(\delta,t)}$ since the signatures are constructed in the diffusion space. In short, the objective of Step 1 is to construct for each data class X_j a signature $SG_j := \{(\boldsymbol{sg}_{j,\ell}, w_{\boldsymbol{sg}_{j,\ell}}) \mid \boldsymbol{sg}_{j,\ell} \in \mathbb{R}^{s(\delta,t)},\ w_{\boldsymbol{sg}_{j,\ell}} \geq 0,\ \ell = 1, \cdots, k_j\}$ that characterizes X_j in the diffusion space $\mathbb{R}^{s(\delta,t)}$.

In Step 1.i we construct the diffusion maps on the union $\cup_{j=1}^m X_j$ in order to embed all data points into the same diffusion coordinate system. The Lafon-Lee's Coarse-Graining algorithm appearing in Step 1.ii is based on the well-known k-means clustering algorithm. (For additional information on k-means algorithm, see e.g., Ref. 9). The main idea is to group the points in Y_j into k_j clusters by minimizing an objective cost functional. In general, the number k_j is determined based on the specific application at hand. The objective functional E is the sum over all clusters of the within-cluster sums of point-to-cluster-centroid squared diffusion distances, more precisely

$$E(\{S_\ell\}_{\ell=1}^{k_j}) := \sum_{\ell=1}^{k_j} \sum_{\boldsymbol{x} \in S_\ell} ||\Psi_t(\boldsymbol{x}) - c(S_\ell)||^2, \tag{1}$$

where S_ℓ denotes the ℓth cluster, $\Psi_t(\boldsymbol{x})$ is the diffusion coordinates of $\boldsymbol{x}$, and $c(S_\ell)$ is the cluster centroid (called *geometric centroid*) of cluster S_ℓ. For a cluster S_ℓ, the *geometric centroid* of S_ℓ is defined as a weighted sum

$$c(S_\ell) := \sum_{\boldsymbol{x} \in S_\ell} \frac{\phi_0(\boldsymbol{x})}{\tilde{\phi}_0(S_\ell)} \Psi_t(\boldsymbol{x}),$$

where $\tilde{\phi}_0(S_\ell) := \sum_{\boldsymbol{x} \in S_\ell} \phi_0(\boldsymbol{x})$, and ϕ_0 is the stationary distribution of the Markov chain as described in Sec. 2.

With Y_j and k_j fixed, minimization of the functional E in Eq. (1) is done via the following steps which guarantee convergence towards a local minimum:

0) Randomly partition the points in Y_j into k_j clusters $\{S_\ell^{(0)}\}_{1 \leq \ell \leq k_j}$ in a uniformly distributed manner,

1) For $p > 0$, update each ℓth partition $(1 \leq \ell \leq k_j)$ according to

$$S_\ell^{(p)} = \{\boldsymbol{x} \in Y_j \mid \ell = \arg\min_i ||\Psi_t(\boldsymbol{x}) - c(S_i^{(p-1)})||^2\}.$$

2) Repeat Step 1 until convergence.

To determine the number k_j, the number of clusters to form for each data set X_j, the common rule of thumb is to apply the so-called Elbow Criterion.[10] For each $k = 1, \cdots, N_j = |X_j|$ (where $|\cdot|$ denotes set cardinality, i.e., $N_j = |X_j|$ is the number of points in X_j), suppose $\{S_\ell^*\}_{\ell=1}^k$ is the minimizer of the functional E in Eq. (1). Let $E_k := E(\{S_\ell^*\}_{\ell=1}^k)$ the total energy of the clustering $\{S_\ell^*\}_{\ell=1}^k$. This number E_k can be viewed as the best clusters-fitness over all sets of k clusters. The best clusters-fitness E_k decreases as k increases, i.e., $\{E_1, \cdots, E_{N_j}\}$ is a decreasing sequence. However, as k increases, the rate of decrease in the sequence $\{E_k\}$ stabilizes, i.e., the 'first derivative' of E_k approaches a constant which equals to 0 in most cases. If we plot $\{E_k\}$ against the number of clusters k, we will see a kink (or elbow) in the plot at $k = k'$ where the rate of decrease in $\{E_k\}$ begins to slow down significantly. The Elbow Criterion says that we choose $k_j = k'$ as the number of clusters to form for X_j. However, in our applications, we modify this condition and choose k_j to be the smallest number at which the rate of decrease in $\{E_k\}$ stabilizes or when the ratio $|E_{k_j} - E_{k_j} - 1|/|E_{k_j+1} - E_{k_j}|$ is the greatest. In Figure 5 we plot the first 11 best clusters-fitness for the experimental data CC2 which we will discuss in more details in our numerical example in Sec. 5.2. According to the Elbow Criterion, the number of clusters to form for this data set is 3. But under our version of the Elbow Criterion, the number of clusters to form is 8.

Suppose $\{S_\ell\}_{\ell=1}^{k_j}$ are the clusters obtained from coarse-graining the set Y_j. We construct the signature SG_j for the data set X_j by setting $\boldsymbol{sg}_{j,\ell} = c(S_\ell)$ and $w_{\boldsymbol{sg}_{j,\ell}} = \frac{|S_\ell|}{|Y_j|}$. In other words, we let the geometric centroids be the cluster representatives and the percentage (or density) of points in the clusters be the weights. For an example, we display in Figure 6 the signature of the shape CC2 to be discussed in Sec. 5.2.

Remarks: (1) Geometric centroids do not necessarily belong to the set of embedded data points Y_j. Therefore, in applications where the cluster representatives are required to be points in Y_j, we may designate the *diffusion centers* to be the cluster representatives. The diffusion center of a cluster S_ℓ, denoted $u(S_\ell)$, is defined to be the point in S_ℓ closest to the geometric centroid $c(S_\ell)$ in the diffusion distance. More precisely,

$$u(S_\ell) := \Psi_t(\boldsymbol{x}^*),\ \boldsymbol{x}^* = \arg\min_{\boldsymbol{x}\in S_\ell} ||\Psi_t(\boldsymbol{x}) - c(S_\ell)||^2.$$

(2) We shall illustrate in our numerical examples that embedding the data points into the diffusion space not only performs dimensionality reduction but also re-arranges the data so that points with similar local geometric properties are close to each other in the diffusion space. Therefore, when we coarse-grain each data set in the diffusion space, points with similar local geometric properties are assembled into the same cluster. Consequently, the clusters forming the signature for a data set can be viewed as a (subsampled) low-dimensional representation of the data that preserves the intrinsic geometry in the data. Furthermore, two different clusters (hence two different cluster representatives) belonging to two different signatures are close in the diffusion space if the local geometry of the points in both clusters are similar, otherwise they are far apart.

To complete Step 2.i in our method, i.e., to compute the Earth Mover's Distance between two signatures SG_{j_1} and SG_{j_2}, we take advantage of the observation described in Remark 2 above and define the cost for transferring one unit of mass from location $\boldsymbol{sg}_{j_1,\ell_1}$ to location $\boldsymbol{sg}_{j_2,\ell_2}$ to be one half of the squared diffusion distance between the two cluster representatives, i.e., $c_{ij} = \frac{1}{2}||\boldsymbol{sg}_{j_1,\ell_1} - \boldsymbol{sg}_{j_2,\ell_2}||^2$. Using squared distances as costs for mass transfer places high penalty on matching points that are far away and favor matching points that are closer. That is, we want the EMD between two data sets to be small if their local geometric properties are similar, otherwise the EMD is large.

5. NUMERICAL EXAMPLES

5.1 Discrimination of sonar signals

In this section we show that our proposed method can be applied to discriminate classes of sonar signals. In Figure 1 we plot three classes of sonar signals. These are recorded waveforms of underwater acoustic near-field scattering experiments. More precisely, we illuminate an aluminum cylindrical casing containing some material inside by sending an acoustic pulse (sinusoid of one period) to the casing and recording the reflected waveforms

from the casing. Each reflected waveform is recorded at 270 time samples. The sampling frequency of the reflected waveforms is set to 500 kHz. The transducer used can generate a pulse (sinusoid) of different duration (frequency). In fact, it generated pulses of 10 kHz to 80 kHz with 2 kHz increment. So, in principle, one should have 36 waveforms per object. Three sets of experiments were conducted. In each set, the same casing with different material inside was used, say, material 1, 2, and 3. Some recordings suffered from severe noise, which were eliminated. For these three data sets, the following frequency sources/recordings were eliminated: 22, 26, 42, 44, 46, 50, 58, 70, and 72 (kHz). Thus, each data set consists of 27 waveforms, each is stored as a vector of length 270.

We treat each waveform as a point in the high-dimensional Euclidean space $\mathbb{R}^{270}$. We then embed all points into a three-dimensional diffusion space, with the parameter ϵ set to be approximately the average of the smallest Euclidean distances between points, i.e.,

$$\epsilon = \frac{1}{N}\sum_{i=1}^{N} \min_{\ell:\boldsymbol{x}_i \neq \boldsymbol{x}_\ell} ||\boldsymbol{x}_i - \boldsymbol{x}_\ell||,$$

where N is the total number of points in all data sets combined. For this example $\epsilon = 0.6$. The time parameter is set to $t = 1$ in all of our examples. Increasing the time parameter t will capture the larger scales in the data, however, in our examples we are interested more in analyzing the data at the local scale which is already set by the value of ϵ. After embedding all waveforms into a three-dimensional diffusion space, we see that each class of waveforms is embedded in a curve (see Figure 2).

Next we construct a signature for each signal class. Since each signal class consists of only 27 data points, coarse-graining in Step 1.ii is unnecessary. We treat each embedded point as a single-point cluster. Thus, the signature of each data class contains 27 clusters, with the cluster representatives being the diffusion coordinates of the data points, i.e., $\boldsymbol{sg}_{j,\ell} = \Psi_1(\boldsymbol{x}_{j,\ell})$ where $\boldsymbol{x}_{j,\ell}$ is the ℓth signal in class j ($j = 1, 2, 3$). Following our method, all clusters shall have the same weight: $w_{\boldsymbol{sg}_{j,\ell}} = \frac{1}{27}$ ($j = 1, 2, 3,\ \ell = 1, \cdots, 27$). Then we define the cost for transferring one unit of mass from location $\boldsymbol{sg}_{j_1,\ell_1}$ to location $\boldsymbol{sg}_{j_2,\ell_2}$ to be one half of the squared diffusion distance between the two points, i.e., $c_{ij} = \frac{1}{2}||\boldsymbol{sg}_{j_1,\ell_1} - \boldsymbol{sg}_{j_2,\ell_2}||^2 = \frac{1}{2}||\Psi_1(\boldsymbol{x}_{j_1,\ell_1})) - \Psi_1(\boldsymbol{x}_{j_2,\ell_2})||^2$. The Earth Mover's Distances (normalized to unit largest distance) between the embedded signal classes are displayed in the left column of Table 1.

Since the waveforms can be treated as points in Euclidean space $\mathbb{R}^{270}$, it is possible for us to measure the EMD between the signal classes without embedding the data in to a diffusion space. This allows us to see the advantage of the diffusion geometry. The weights and costs are defined in the same way as above. The (normalized) EMD between the signal classes without the embedding step are displayed in the right column of Table 1. The difference in the EMD values from the two experiments shows that embedding the waveforms into a diffusion space helps to distinguish the difference between all three signal classes. Without embedding into a diffusion space, the EMD between class 1 and class 2 is very close to that between class 1 and class 3.

5.2 Quantification of shape deformations

In this example, we show that our proposed method can be applied to quantify the difference between two shapes. Our method can also be applied in the same manner to quantify how much one shape has evolved or deformed from another. In this example, each shape is represented by its boundary which is given as a set of points in $\mathbb{R}^2$. In Figure 3 we plot the segmented boundaries of three sagittal-viewed MRI scans of the corpus callosum structure in the brain. The number of data points in shape CC1, CC2 and CC3 are 656, 649 and 630, respectively.

We embed all data points into a two-dimensional diffusion space, then coarse-grain to reduce the dimensionality in the data. The number of clusters k_j to form for each data set is determined by our version of the Elbow Criterion, with an additional constraint that $0 < k_j \leq 20$. Signatures for each shape are then constructed using geometric centroids and percentage of point densities as cluster representatives and weights. The cost of mass transfer is again defined as one half of the squared diffusion distance between any two cluster representatives.

In this example, a careful selection of the parameter ϵ is required. In general, any closed curve in $\mathbb{R}^2$ is embedded onto a circle in any diffusion space. However, when ϵ is small enough, the spectrum decay is stagnated,

and the eigenvectors corresponding to the top two eigenvalues starts to deviate from sine and cosine functions. So if the curves are different, the diffusion coordinates will show some differences. In Figure 4, we plot the diffusion coordinates for the three corpus callosum shapes. The value for ϵ is 0.7. Observe that the embeddings corresponding to CC1 and CC3 are very similar but not exactly the same while the one corresponding to CC2 is very different from the others.

The advantage of embedding the data points into a diffusion space before clustering them to form signatures is highlighted in this example. In the diffusion space, closeness is determined by the local geometry in the data. In this example, two embedded points are close in the diffusion space if they are close along the curve in the ambient space, instead of being close in terms of the ambient Euclidean distance. To illuminate this aspect, we plot in Figure 7 the results of coarse-graining in the diffusion space and standard k-means clustering in the ambient space. The points of shape CC3 are grouped into six clusters by each algorithm. Recall that coarse-graining is performed in the diffusion space. The plot on the left of Figure 7 displays the clusters of points in the ambient space corresponding to the clustering results from coarse-graining in the diffusion space. Each cluster is displayed at a different gray level. We can see that subsampling in the diffusion space preserves the local geometry of the original data, i.e., the ordering of points along the curve is preserved. Comparing this to the result of the standard k-means clustering, we see that the clusters are completely determined by the ambient Euclidean distance, not by the geometry of the curve. As a consequence, if we construct signatures for the data in the ambient space, the clusters that we would obtain carry no geometric information of the data. And thus, we would not be able to say that the EMD between two signatures indicates how different the two curves are geometrically, as is the case when the signatures are constructed via coarse-graining in the diffusion space.

One drawback in the proposed method is the nonuniqueness of the solution of the Coarse-Graining algorithm. The clustering result depends on the clusters initialization at Step 0 in the computation for a minimum of the functional E in Eq. (1). If we apply the Coarse-Graining algorithm twice on the same set of data, we may obtain different clustering results, hence we get two different signatures for the same data set. Consequently, if we repeat our proposed method on the same sets of data the second time, the EMD values between the data sets may change. However, although the EMD values may change, the discriminative implications from these values do not change. For an example, we repeated applying our method on the three shapes CC1, CC2, and CC3 one hundred times. Each time we collect a set of three EMD values (the EMD between the three pairs of shapes: EMD(CC1,CC2), EMD(CC1,CC3) and EMD(CC2,CC3)). And every time, the EMD values imply that the shapes CC1 and CC3 are very similar, but not the same, and the shape CC2 is very different from both shapes CC1 and CC3. We have plotted the 100 sets of EMD values in Figure 8. One specific set of three EMD values (normalized to unit largest value) is EMD(CC1,CC2) = 0.9958, EMD(CC1,CC3) = 0.0085, EMD(CC2,CC3) = 1. Overall 100 sets, the average of these values are AVG_EMD(CC1,CC2) = 0.9971, AVG_EMD(CC1,CC3) = 0.0089, AVG_EMD(CC2,CC3) = 1.

We end this section with a final remark. The proposed method is not invariant under scaling, rotation or translation. For shape comparison applications, one could apply first as a preprocessing step some rigid registration to correct any variations in rotation and translation. On the other hand, detection of scalings are desirable in many applications. For example, any shrinking or growth of the corpus callosum should be detected in the diagnosis of agenesis of the corpus callosum. Therefore, the proposed method is applicable in this case. However, if one's problem should be scale invariant, then the proposed method should not be considered.

6. CONCLUSIONS

We have proposed a novel application of diffusion maps combined with Earth Mover's Distance to produce a dissimilarity measure for discrimination of various different classes of data. In our examples, the proposed method successfully distinguished different classes of acoustic waveforms corresponding to reflected sinusoidal signals targeted at different types of underwater objects. This promises applicability of our proposed method to detection of underwater mines via classification of sonar signals. We have also shown that our proposed method can be applied to quantify deformations of shapes, although with some restrictions. Such applications are extremely useful in medical image analysis where deformations or changes of a certain brain structure reveal much about certain medical conditions of the patient.

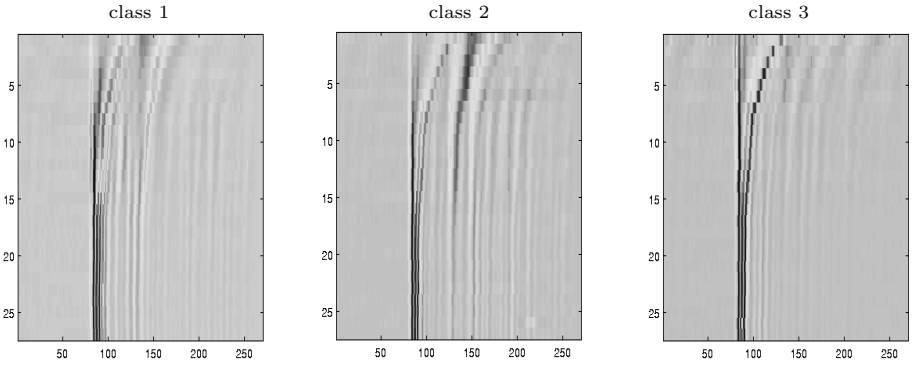

Figure 1. Sonar signal recordings of sinusoidal reflections from three different underwater objects. The vertical axis corresponds to sampling frequencies and the horizontal axis corresponds to recording time.

ACKNOWLEDGMENTS

We would like to thank Naval Surface Warfare Center, Panama City, FL., for providing us with the acoustic scattering data. This research was partially supported by NSF grants DMS-0135345, DMS-0410406, and ONR grant N00014-07-1-0166.

REFERENCES

1. R. R. Coifman and S. Lafon, "Diffusion Map", Appl. Comput. Harmon. Anal., **21**, pp. 5-30, 2006.
2. S. Lafon, "Diffusion Maps and Geometric Harmonics", Ph.D. Dissertation, Yale University, May 2004.
3. S. Lafon and A. B. Lee, "Diffusion maps and coarse-graining: A unified framework for dimensionality reduction, graph partitioning and data set parameterization", IEEE Trans. Pattern Analysis and Machine Intelligence, **28**(9), pp. 1393-1403, 2006.
4. M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation", Neural Computation, **15**(6), pp. 1373-1396, 2003.
5. Y. Rubner and C. Tomasi, *Perceptual Metrics for Image Database Navigation*, Kluwer Academic Publishers, Boston, 1999.
6. Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth mover's distance as a metric for image retrieval", IJCV, **40**(2), pp. 99-121, 2000.
7. E. Levina and P. Bickel, "The Earth Mover's Distance is the Mallows Distance: Some insights from statistics", *Computer Vision 2001, ICCV 2001 Proceedings, Eighth IEEE International Conference*, 2, pp. 251-256, 2001.
8. L. C. Evans, "Partial Differential Equations and Monge-Kantorovich Mass Transfer" (lecture notes), `www.math.berkeley.edu/~evans/Monge-Kantorovich.survey.pdf`.
9. J. Kogan, *Introduction to Clustering Large and High-Dimensional Data*, Cambridge Univ. Press, 2007.
10. T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Section 14.3.11, Springer, 2001.

Table 1. EMD between signal classes with waveforms embedded into a diffusion space (left) and between signal classes with waveforms treated as points in $\mathbb{R}^{270}$ (right). In each case, the EMD values are normalized so that the largest of the three values is 1.

	class 1	class 2	class 3
class 1	0	1	0.57
class 2		0	0.88
class 3			0

	class 1	class 2	class 3
class 1	0	1	0.83
class 2		0	0.87
class 3			0

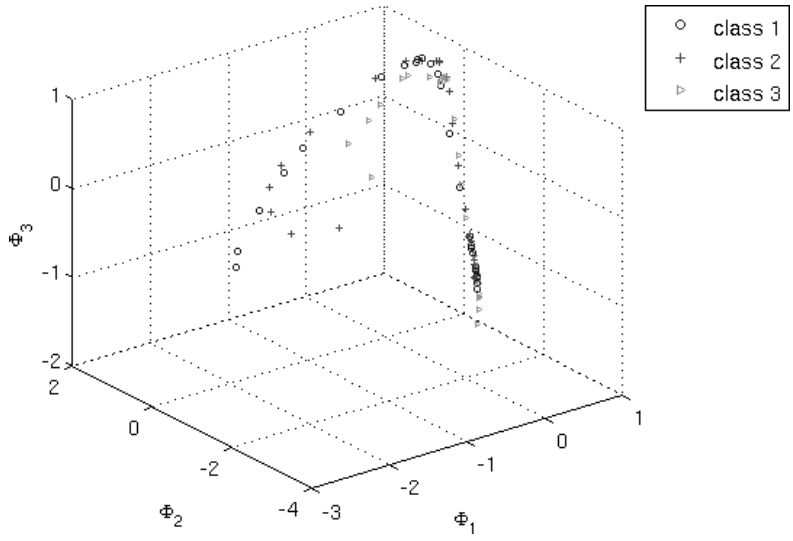

Figure 2. Embedding of all waveforms into three-dimensional diffusion space.

Figure 3. Three corpus callosum shapes segmented from sagittal view of three MRI scans. The shapes are overlaid to portray differences between them.

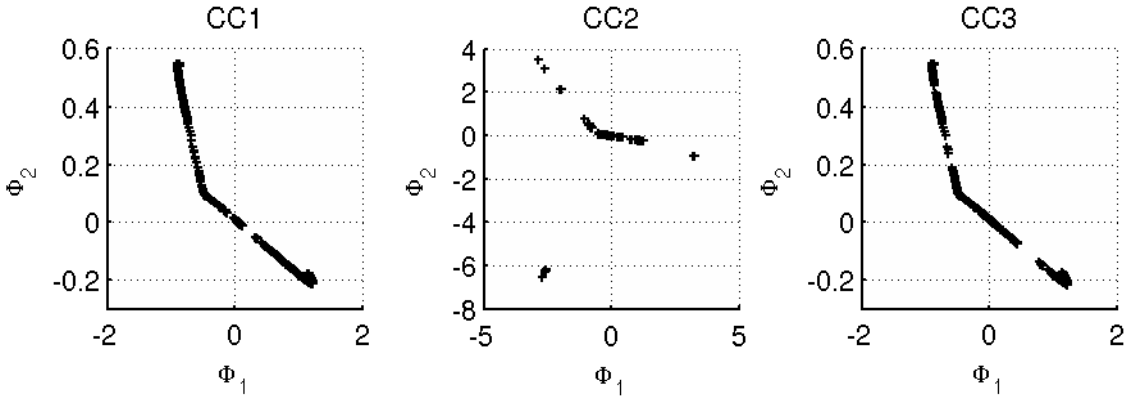

Figure 4. Diffusion coordinates of the three sets of points describing the three corpus callosum shapes.

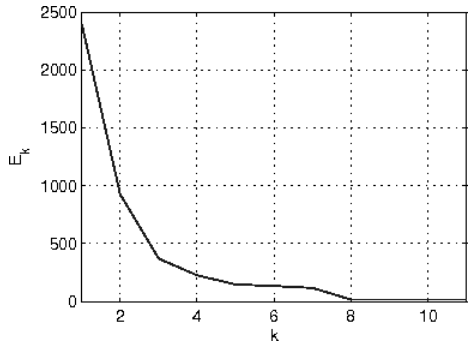

Figure 5. Best clusters-fitness as a function of number of clusters k. The data set used to generate the values E_k is shape CC2. According to the Elbow Criterion, the number of clusters to form for shape CC2 is 3. According to our version of the Elbow Criterion, the number of clusters to form is 8.

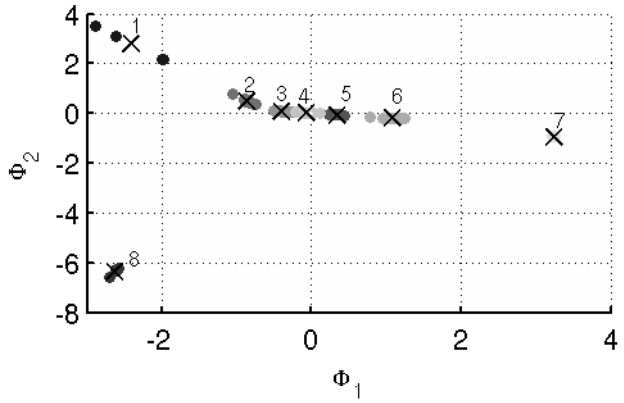

Figure 6. Signature of shape CC2 in diffusion space. This signature has eight clusters, each displayed at a different gray-level. The cluster representatives (which are the same as the geometric centroids) are indicated by an 'x'. Each cluster is numbered from 1 to 8. The signature's cluster representatives and weights are $SG = \{((-2.4, 2.8), \frac{16}{649}), ((-0.9, 0.5), \frac{216}{649}), ((-0.4, 0.08), \frac{65}{640}), ((-0.07, 0.02), \frac{78}{649}), ((0.3, -0.05), \frac{52}{649}), ((1.1, -0.2), \frac{167}{649}), ((3.2, -0.9), \frac{24}{649}), ((-2.6, -6.4), \frac{31}{649})\}$.

Figure 7. Result of coarse-graining in the diffusion coordinate system (left) and result of standard k-means clustering in the ambient space (right). Points of shape CC3 are grouped into six clusters by both algorithms. Different clusters are displayed at different gray-level intensities. **Left:** clustering results of coarse-graining in the diffusion coordinate system. Each cluster consists of adjacent points forming a segment on the original closed curve. The boundary of each segment is marked by a white dot, so there are two boundary points per cluster. **Right:** clustering results of standard k-means in the original coordinate system. The boundary points of each cluster are displayed as white dots. Four of the six clusters are consisted of two parallel segments, so there are four boundary points in each of these four clusters.

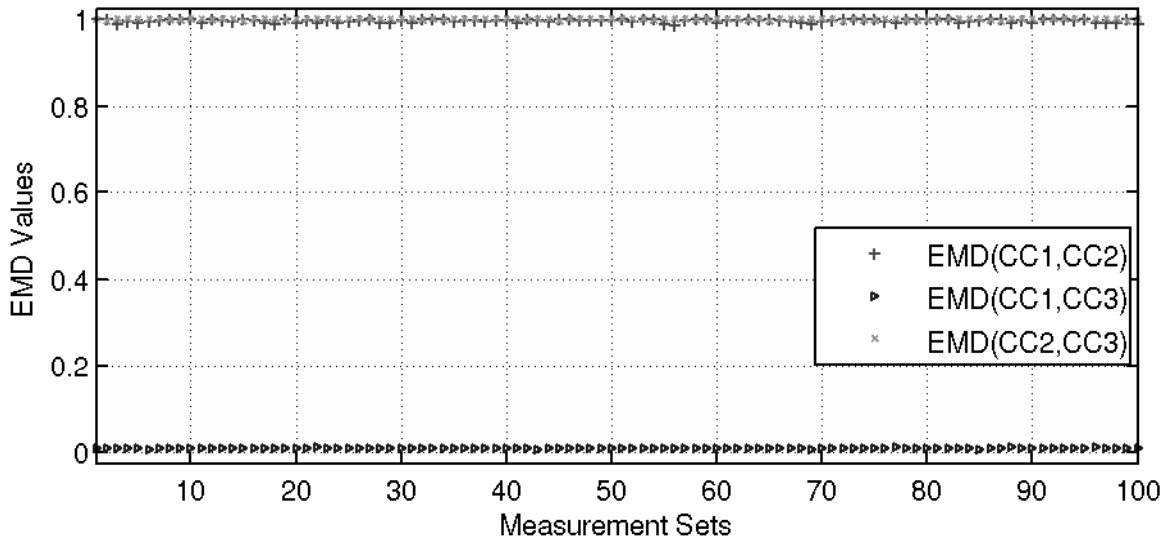

Figure 8. Plot of 100 sets of Earth Mover's Distances between the three corpus callosum shapes. Each set consists of three EMD values: EMD(CC1,CC2), EMD(CC1,CC3), and EMD(CC2,CC3). The horizontal axis is the set number and the vertical axis are the EMD values. The EMD between each pair of shapes is plotted using a different marker.