Multidimensional Artificial Field Embedding with Spatial Sensitivity

Dalton Lunga, *Student member, IEEE,* and Okan Ersoy, *Fellow, IEEE*

***Abstract*—Multidimensional embedding is a technique useful for characterizing spectral signature relations in hyperspectral images. However, such images consist of disjoint similar spectral classes that are spatially sensitive, thus presenting challenges to existing graph embedding tools. Robust parameter estimation is often difficult when the image pixels contain several hundreds of bands. In addition, finding a corresponding high quality lower dimensional coordinate system to map signature relations remains an open research question. We answer positively on these challenges by first proposing a combined kernel function of spatial and spectral information in computing neighborhood graphs. We further adapt a force field intuition from mechanics to develop a unifying nonlinear graph embedding framework. The generalized framework leads to novel unsupervised multi-dimensional artificial field embedding techniques that rely on the simple additive assumption of pair-dependent attraction and repulsion functions. The formulations capture long range and short range distance related effects often associated with living organisms and help to establish algorithmic properties that mimic mutual behavior for the purpose of dimensionality reduction. In its application, the framework reveals strong relations to existing embedding techniques, and also highlights sources of weaknesses in such techniques. As part of evaluation, visualization, gradient field trajectories, and semisupervised classification experiments are conducted for image scenes acquired by multiple sensors at various spatial resolutions over different types of objects. The results demonstrate the superiority of the proposed embedding framework over various widely used methods.**

I. Introduction

Airbone and space-based sensors continue to enable greater improvement in quality image acquisition with the goal of providing detailed information for material identification. The images are characterized by high spectral resolution which generates several hundred of bands capturing the electromagnetic reflectance properties of different materials. However, having many bands (or channels) poses several challenges to conventional land cover studies that include classification algorithms due to the curse of dimensionality. Other additional challenges include the inherent nonlinear characteristics that stem from bidirectional reflectance distribution function(BRDF) [1]. BRDF often leads to variations in spectral reflectance of different classes as a function of position in the landscape, depending on the local topology. Furthermore, for imagery that is acquired over coastal environments (*e.g.* coastal wetlands), the variable presence of water in pixels as a function of position in landscape presents more sources of nonlinearities. Recent research efforts have identified dimensionality reduction and manifold learning techniques as key for preprocessing hyperspectral images.

Notwithstanding individual differences in efficiency, accuracy, and application to hyperspectral images, dimensionality reduction methods share some features including better compression, better visualization, and extraction of useful classifier input features. Widely used techniques include linear based formulations that are easy to implement, *e.g.* principal component analysis(`PCA`) [2], the multidimensional scaling(`MDS`) [3], the local Fisher discriminant analysis (`LFDA`) [4], and the local Fisher discriminant analysis (`SELF`) [5]. Further developments in hyperspectral sensing have enabled the acquisition of greater details about objects on the earth surface which poses a challenge for linear dimensionality reduction techniques. Better information extraction can be accomplished by employing nonlinear methods such as the maximum variance unfolding (`MVU`) [6]- a method that computes maximum variance embedding maps subject to preserving local distances, the locally linear embedding(`LLE`) [7] - a method that represents the relations of each neighborhood by linear coefficients that best reconstruct each data point from its neighbors, and the laplacian eigenmaps Laplacian(`LE`) [15], which draws on the correspondence between the graph Laplacian, the Laplace Beltrami operator on a manifold, and the connections to the heat equation, to devise a geometrically motivated algorithm for constructing a representation for data sampled from a low m-dimensional manifold embedded in a higher d-dimensional space. Both `LLE`, and `LE` solution spaces consist of the trailing eigenvectors obtained by the eigendecomposition of the transformed coefficient and the Laplacian matrices, respectively. The assumption of linearity on the local neighborhoods of data and their piecewise combination to form global nonlinear structures was used to propose isometric feature mapping(`Isomap`) [8]. As in `LLE`, `Isomap` takes as input a high dimensional neighborhood graph whose edge weights are computed from pairwise distances to characterize object similarities. It then obtains a globally optimal coordinate system for the nonlinear data through an MDS solution. Recently, probabilistic methods have also been proposed. These include the stochastic neighbor embedding(`SNE`) [12], a method that represents each object by a mixture of widely separated low dimensional factors capturing some of the local structure, and establishes global formations of clusters of similar maps. The method called student t-distribution based stochastic neighbor embedding(`tSNE`) [11] has been proposed as a variant of `SNE` that assumes the probability relations in the lower dimensional

D. Lunga is with the Department of Electrical and Computer Engineering, Purdue University. He is also affiliated with the CSIR-Meraka Institute, Brummeria, Pretoria, South Africa e-mail: dlunga@csir.co.za.

O. Ersoy is with the Department of Electrical and Computer Engineering, Purdue University, West Lafayette, IN, 47907-0501, USA e-mail: ersoy@purdue-edu.

space to be inversely proportional to the distance between pairs of maps.

Each of these techniques represents an attempt to search for a coordinate representation that resides on the nonlinear data manifold. Very few of the existing methods have been successfully used in the analysis of hyperspectral data. For example, the hybrid isometric mapping (Isomap) and locally linear embedding(LLE), were both combined to develop an algorithm that can handle the compression and classification of large sample images [1]. More recently, a comparative study was conducted to evaluate the effect of various nonlinear manifold learning algorithms when reducing the dimension of hyperspectral data [13]. The results obtained from these studies highlight the performance improvement on classification tasks. However, they also strongly suggest the inability of existing dimensionality reduction methods to take advantage of the disjoint class structure that exists in hyperspectral imagery. Embedding algorithms with excessive inability to discriminate dissimilar objects (or handle disjoint classes) can be characterized as suffering from the *crowding problem* [11]. Within the hyperspectral context, the crowding problem can be defined as the tendency of collapsing pixel maps towards the center of the embedding space. This phenomenon causes embedding algorithms to fail to establish discriminative boundaries that are required for improved classification accuracy.

In this study, we show that the crowding of pixels can be handled in two phases. We first propose a general embedding framework that can be used as a platform for developing new dimensionality reduction models. Secondly, we show that the disjoint nature often present in hyperspectral images is key to mitigating the crowding problem and can be encoded onto the neighborhood graph through a combined spectral-spatial kernel function.

The new dimensionality reduction framework is presented as the multidimensional artificial field embedding(MAFE), with an optimization scheme that aims to establish a minimum energy configuration state of the high dimensional neighborhood graph. The framework draws on the force field intuition from mechanics, whereby pair-dependent attraction and repulsion functions are designed to reflect long range and short range force effects. The functions are superposed to generate an odd-function that invokes a pairwise interaction force between pairs of maps in the embedding space. The framework benefits from the design of simple embedding algorithms whose objective functions are easier to differentiate. As a second contribution, we present a novel spatially-sensitive kernel function for computing high dimensional neighborhood graphs. The kernel function encodes spatial details that are essential for maintaining disjoint spectral classes in hyperspectral embedding applications. The final embedding algorithm enables lower dimensional pixel maps with high similar values to form segmented regions that are sensitive to spatial details. Furthermore, we identify links to existing approaches that includes SNE [12], tSNE [11], and the recent spherical stochastic neighbor embedding(sSNE) [19], all methods exhibiting powerful data visualization capabilities.

This paper is organized as follows: Section II provides a brief review of the force field intuition, and subsequently presents a general dynamic system based graph embedding approach. Section III describes reformulations of some existing methods within the dynamic system framework. Section IV describes the functional forms used in formulating the new multidimensional artificial field embedding - bounded repulsion(MAFE−BR) model. Section V describes the bilateral kernel function for inducing spatial-sensitivity to the neighborhood graph. The optimization algorithm and its properties are presented in Section VI. Experimental results are summarized in Section VII. Section VIII provides a discussion and future work. Conclusions are presented in Section IX.

II. MAFE Formulation

Force field formulations are widely used in robotics studies, *e.g.* [20]–[22] in which related models are described to study the stability and motion planning of robots, respectively. Here we make an analogy of objects moving in a coordinate space and apply this framework to devise a general approach to graph based embedding. MAFE imagines each image pixel to be associated with a vertex of a high dimensional neighborhood graph that has a corresponding optimal map in the lower dimensional space. In each map, we further consider a particle in motion whose movement determines the position of the embedding map. Each embedding map has a dual role - both attractive and repulsive, that is dynamically determined by the changing distance based interactions during the optimization. We treat the dimensionality reduction problem as a task that requires solving N unconstrained optimization problems to obtain the minimum-energy configuration state which yields the graph topology that preserve neighborhood relations.

A. MAFE *General Graph Embedding Framework*

Let $\mathcal{G} = (\mathcal{E}, \mathcal{V})$ be a finite undirected graph with vertices $\mathcal{V}$, edges $\mathcal{E}$ and no self loops. We designate elements of $\mathcal{E}$ as ideal springs. Let $\mathcal{S} = \{(w_{ij}, k_{ij})\}$ be the spring properties between each vertex i and j for all $\{v_i, v_j\} \in \mathcal{E}$, where w_{ij} is the normalized length without compression or extension computed for each observed pair in $\boldsymbol{Y} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_N\}$, with $\boldsymbol{y}_i \in \mathbb{R}^d$, and $k_{ij} = 1$ is the force constant. A graph with relation $\mathcal{S}$ is called a neighborhood spring graph and we can denote it by $\mathcal{G}_{\mathcal{S}}$.

An embedding of $\mathcal{G}_{\mathcal{S}}$ is an assignment of vertices into a m-dimensional Euclidean space $\mathbb{R}^m$ (*i.e.* m describes a lower dimensional space). Let $\boldsymbol{Z} = \{\boldsymbol{z}_1, \boldsymbol{z}_2, \cdots, \boldsymbol{z}_N\}^T$ be the assigned embedding of $\mathcal{G}_{\mathcal{S}}$, where $\boldsymbol{z}_i \in \mathbb{R}^m$ is the position of vertex i's map. When framed as a graph embedding task, where on each vertex we imagine a particle and the edges as representing spring force laws, the problem of dimensionality reduction simply becomes that of establishing a minimum energy embedding that is governed by the structure in $\boldsymbol{W} = [w_{ij}]$. We hope such an embedding yields the maps that preserve pairwise distances described by the neighborhood graph $\mathcal{G}_{\mathcal{S}}$.

Finding such a mapping is at the heart of every dimension reduction model, and it is the subject that we discuss next. We first give a mechanics interpretation of the graph embedding framework as follows: imagine the existence of a

particle on every $\boldsymbol{z}_i \in \boldsymbol{Z}$, that is moving with the velocity of $\boldsymbol{Z}$'s centroid. With the following change of notation to denote the embedding positions as a *state* of a graph, we let $\mathcal{Z} = \left\{\boldsymbol{z}_1^T, \boldsymbol{z}_2^T, \cdots, \boldsymbol{z}_N^T\right\}^T$ be a long vector in $\mathbb{R}^{Nm}$. Thus, we only consider the motion dynamics of individual maps, not the motion of the group. We assume that all individual maps move simultaneously, and each map i is aware of the position and the strength of forces that exist with positions of all other vertices. The positions, $\boldsymbol{z}_i$'s, of individuals relative to the group centroid can change through the rearrangement due to pair-dependent interactions. Assuming such motion is to change in a continuous time, the velocity as determined by the effect of group members on each vertex i, at position $\boldsymbol{z}_i$ is described by

$$\dot{\boldsymbol{z}}_i = \sum_{j \neq i} F^{ij}(\boldsymbol{z}_i - \boldsymbol{z}_j), \quad i = 1, \cdots, N \tag{1}$$

where $F^{ij}(\boldsymbol{z}_i - \boldsymbol{z}_j) = (\boldsymbol{z}_i - \boldsymbol{z}_j)\left\{F_r^{ij}(\|\boldsymbol{z}_i - \boldsymbol{z}_j\|) - F_a^{ij}(\|\boldsymbol{z}_i - \boldsymbol{z}_j\|)\right\}$ describes pairwise symmetric interactions between the ith and jth maps. Symmetry of the function follows from the fact that if map i is attracted to map j, then j is attracted to i. $F_r^{ij} : \mathbb{R}^+ \to \mathbb{R}^+$ denotes the magnitude of the repulsion term, whereas $F_a^{ij} : \mathbb{R}^+ \to \mathbb{R}^+$ represents the magnitude of the attraction term.

B. Embedding Force Fields Function Properties

We assume that at large distances, the attraction dominates, and at short distances, the repulsion dominates while in between there is a unique distance at which the attraction and the repulsion balance. A suitable candidate for $F^{ij}(\boldsymbol{z}_i - \boldsymbol{z}_j)$ should obey the following *embedding force field* properties:

1) There is a pair-equilibrium distance ϵ_{ij} at which $F_r^{ij}(\epsilon_{ij}) = F_a^{ij}(\epsilon_{ij})$, else $F_a^{ij}(\|\boldsymbol{z}_i - \boldsymbol{z}_j\|) > F_r^{ij}(\|\boldsymbol{z}_i - \boldsymbol{z}_j\|)$ for $\|\boldsymbol{z}_i - \boldsymbol{z}_j\| > \epsilon_{ij}$ or $F_a^{ij}(\|\boldsymbol{z}_i - \boldsymbol{z}_j\|) < F_r^{ij}(\|\boldsymbol{z}_i - \boldsymbol{z}_j\|)$ for $\|\boldsymbol{z}_i - \boldsymbol{z}_j\| < \epsilon_{ij}$.
2) F^{ij} is an odd function, *i.e.* $F^{ij}(-(\boldsymbol{z}_i - \boldsymbol{z}_j)) = -F^{ij}(\boldsymbol{z}_i - \boldsymbol{z}_j)$, therefore symmetric with respect to the origin.
3) There exist pair dependent functions $U_{att}^{ij} \to \mathbb{R}^+ \to \mathbb{R}^+$ and $U_{rep}^{ij} \to \mathbb{R}^+ \to \mathbb{R}^+$ such that

$$\nabla_{\boldsymbol{z}_i} U_{att}^{ij}(\|\boldsymbol{z}_i - \boldsymbol{z}_j\|) = F_a^{ij}(\|\boldsymbol{z}_i - \boldsymbol{z}_j\|)(\boldsymbol{z}_i - \boldsymbol{z}_j)$$
$$\nabla_{\boldsymbol{z}_i} U_{rep}^{ij}(\|\boldsymbol{z}_i - \boldsymbol{z}_j\|) = F_r^{ij}(\|\boldsymbol{z}_i - \boldsymbol{z}_j\|)(\boldsymbol{z}_i - \boldsymbol{z}_j)$$

U_{att}^{ij} and U_{rep}^{ij} are viewed as *artificial* attraction and repulsion potential energies. The combined term $(\boldsymbol{z}_i - \boldsymbol{z}_j)F_r^{ij}(\|\boldsymbol{z}_i - \boldsymbol{z}_j\|)$ represents the actual repulsion, whereas the term $-(\boldsymbol{z}_i - \boldsymbol{z}_j)F_a^{ij}(\|\boldsymbol{z}_i - \boldsymbol{z}_j\|)$ represents the actual attraction. The vector $(\boldsymbol{z}_i - \boldsymbol{z}_j)$ establishes the alignment for the interaction forces to act along opposing directions. The functions describe the reactive approach by potential fields in which the trajectories of the particles motion are not planned explicitly. Instead the interactions of every map with its neighbors is a *superposition* of fields that enable its position to cope with the changing environment of other maps. We can rewrite the motion dynamics to reflect the resultant forces on each individual map as

$$\dot{\boldsymbol{z}}_i = -\sum_{j \neq i} \left\{\nabla_{\boldsymbol{z}_i} U_{att}^{ij}(\|\boldsymbol{z}_i - \boldsymbol{z}_j\|) - \nabla_{\boldsymbol{z}_i} U_{rep}^{ij}(\|\boldsymbol{z}_i - \boldsymbol{z}_j\|)\right\}$$

The assumption of each map moving along the negative gradient implies that, to achieve a minimum-energy configuration of a graph $\mathcal{G}_{\mathcal{S}}$, we should choose the attraction and repulsion potentials such that the minimum of $U_{att}^{ij}(\|\boldsymbol{z}_i - \boldsymbol{z}_j\|)$ occurs on or around $\|\boldsymbol{z}_i - \boldsymbol{z}_j\| = 0$, whereas the minimum of $-U_{rep}^{ij}(\|\boldsymbol{z}_i - \boldsymbol{z}_j\|)$ (or maximum of $U_{rep}^{ij}(\|\boldsymbol{z}_i - \boldsymbol{z}_j\|)$) occurs on or around $\|\boldsymbol{z}_i - \boldsymbol{z}_j\| \to \infty$, and that the minimum of the combination $U_{att}^{ij}(\|\boldsymbol{z}_i - \boldsymbol{z}_j\|) - U_{rep}^{ij}(\|\boldsymbol{z}_i - \boldsymbol{z}_j\|)$ occurs at $\|\boldsymbol{z}_i - \boldsymbol{z}_j\| = \epsilon_{ij}$.

The above framework can be written to represent the reactive potentials on each individual map i as

$$U_i(\mathcal{Z}) = \sum_{j \neq i} \left\{U_{att}^{ij}(\|\boldsymbol{z}_i - \boldsymbol{z}_j\|) - U_{rep}^{ij}(\|\boldsymbol{z}_i - \boldsymbol{z}_j\|)\right\} \tag{2}$$

while the total superposed potential function on $\mathcal{G}_{\mathcal{S}}$ is defined by

$$U(\mathcal{Z}) = \sum_{i=1}^{N} U_i(\mathcal{Z}) \tag{3}$$

By adapting the above functional properties, new embedding models can simply be derived by solving the optimization problem of the form

$$\mathcal{Z}^\star = \underset{\mathcal{Z} \in \mathbb{R}^{Nm}}{\operatorname{argmin}} U(\mathcal{Z}) \tag{4}$$

With some parameter adjustments on $F^{ij}(\cdot)$, the embedding maps will thus converge to a minimum-energy configuration that yields the required dimension reduced maps. The minimum-energy configuration state is described by $\mathcal{Z}^\star$. This state defines the central embedding points $\boldsymbol{Z}^\star$ where the pairwise repulsion and attraction forces balance. The following section makes brief connections with existing techniques to highlight the general nature of MAFE's framework.

III. MAFE Connections To Existing Methods

In this section, we present a MAFE interpretation of stochastic neighbor embedding(SNE) [12] and the student-t stochastic neighbor embedding([11]). The MAFE interpretation of the spherical stochastic neighbor embedding(sSNE) [19] method follows the same approach.

A. Stochastic Neighbor Embedding

Stochastic neighbor embedding ([12]) is a method for preserving probabilities on lower dimensional manifolds that are nonlinear. SNE assumes that edge weights are antisymmetric probabilities w_{ij} (i.e. $w_{ij} \neq w_{ji}$) of pairs of vertices being neighbors in the higher dimensional space. However, our presentation focuses on the symmetric version where $w_{ij} = w_{ji}$ for all pairs of vertices. The high dimensional edge weights are defined using the Gaussian functions of the form

$$w_{ij} = \frac{\exp\{-\frac{\|\boldsymbol{y}_i - \boldsymbol{y}_j\|^2}{2\sigma_i}\}}{\sum_{r=1, r\neq i} \exp\{-\frac{\|\boldsymbol{y}_r - \boldsymbol{y}_i\|^2}{2\sigma_i}\}} \tag{5}$$

where σ_i is computed using a binary search method ensuring that the entropy of the distribution P_i is approximately $\log(k)$, with k being the effective number of neighbors. In the lower dimensional space, we also assume symmetric Gaussian probabilities $\hat{w}_{ij}$ (i.e. $\hat{w}_{ij} = \hat{w}_{ji}$), between observation maps. Therefore the embedding graph weights are computed as

$$\hat{w}_{ij} = \frac{\exp\{-\|\boldsymbol{z}_i - \boldsymbol{z}_j\|^2\}}{\sum_{r=1,r\neq i}\exp\{-\|\boldsymbol{z}_r - \boldsymbol{z}_i\|^2\}} \tag{6}$$

Each $\boldsymbol{z}_i \in \mathbb{R}^m$ is the corresponding lower dimensional map of the observation $\boldsymbol{y}_i \in \mathbb{R}^d$. SNE proceeds to compute for the maps by minimizing a sum of Kullback Leibler(KL) objective functions

$$\sum_i KL(P_i||Q_i) = \sum_i \sum_{j\neq i} w_{ij}\log(\frac{w_{ij}}{\hat{w}_{ij}}) \tag{7}$$

The goal of (7) is to minimize the distortion between each of the N high dimensional neighborhood distributions P_i's and their corresponding lower dimensional neighborhood distributions Q_i's. Embedding results obtained from this approach have so far proven to be superior when compared to methods that include locally linear embedding(LLE) [7], MDS [3], and Isomap [8]. However, the optimization algorithm is very unstable, and leads to a lot of experimentally defined parameters for attaining meaningful results. A further expansion on (7) while ignoring terms that do not depend on $\hat{w}_{ij}$, reveals the *log-sum* term as a source of difficulty when computing the gradient and increases the nonlinearity of the model. The expansion of (7) leads to

$$\begin{aligned} U^{\texttt{SNE}} &= -\sum_{i,j\neq i} w_{ij}\log\hat{w}_{ij} \\ &= \sum_{i,j\neq i} w_{ij}\|\boldsymbol{z}_i - \boldsymbol{z}_j\|^2 + \log\sum_{r\neq i}\exp\{-\|\boldsymbol{z}_r - \boldsymbol{z}_i\|^2\} \end{aligned}$$

Computing the negative gradient yields the maps motion dynamics equation

$$\dot{\boldsymbol{z}}_i^{\texttt{SNE}} = -4\sum_{j\neq i}(\boldsymbol{z}_i - \boldsymbol{z}_j)\left\{w_{ij} - \frac{\exp\{-\|\boldsymbol{z}_i - \boldsymbol{z}_j\|^2\}}{\sum_{r\neq i}\exp\{-\|\boldsymbol{z}_r - \boldsymbol{z}_i\|^2\}}\right\}$$

where (8) describes a superposition of two potential energy functions whose gradients are given by (8). We identify the attractive force, $-w_{ij}(\boldsymbol{z}_i - \boldsymbol{z}_j)$, and a repulsion force $\frac{(\boldsymbol{z}_i - \boldsymbol{z}_j)\exp\{-\|\boldsymbol{z}_i - \boldsymbol{z}_j\|^2\}}{\sum_{r=1,r\neq i}\exp\{-\|\boldsymbol{z}_r - \boldsymbol{z}_i\|^2\}}$. The interpretation of (8) follows the intuition of MAFE, that is at longer distances, the embedding maps start to form clusters as determined by the attraction forces, while the repulsion forces are very negligible. As $\|\boldsymbol{z}_r - \boldsymbol{z}_i\| \to 0$ for each (i, r) pair, the repulsion magnitude dominates the interaction force vector. The convergence of the algorithm is established when the forces balance.

B. t-Stochastic Neighbor Embedding

t-Stochastic Neighbor Embedding([11]) is similar to SNE except that the lower dimensional maps are assumed to be better modeled by a *Student t-distribution* of degree one. This simple modification leads to a complete improvement of results over SNE. The improvement is due to the pair-dependent inverse distance relation introduced by the Student t-distribution. As such $\hat{w}_{ij}$ is defined as

$$\hat{w}_{ij} = \frac{(1 + \|\boldsymbol{z}_i - \boldsymbol{z}_j\|^2)^{-1}}{\sum_{r=1,r\neq i}(1 + \|\boldsymbol{z}_r - \boldsymbol{z}_i\|^2)^{-1}} \tag{8}$$

tSNE formulates its cost function as in (7) and proceed to compute for the maps by minimizing a sum of Kullback Leibler(KL) objective functions

$$\sum_i^N KL(P_i||Q_i) = \sum_i \sum_{j\neq i} w_{ij}\log(\frac{w_{ij}}{\hat{w}_{ij}}) \tag{9}$$

Computing the negative gradient yields

$$\dot{\boldsymbol{z}}_i^{\texttt{tSNE}} = -4\sum_{j\neq i}\left\{\frac{w_{ij}(\boldsymbol{z}_i - \boldsymbol{z}_j)}{1 + \|\boldsymbol{z}_i - \boldsymbol{z}_j\|^2} - \frac{\frac{(\boldsymbol{z}_i - \boldsymbol{z}_j)}{(1+\|\boldsymbol{z}_i - \boldsymbol{z}_j\|^2)^2}}{\sum_{r\neq i}\frac{1}{(1+\|\boldsymbol{z}_r - \boldsymbol{z}_i\|^2)}}\right\}$$

This derivation demonstrates that tSNE is a special case MAFE model. We can identify the attractive force $-(\boldsymbol{z}_i - \boldsymbol{z}_j)\frac{w_{ij}}{1+\|\boldsymbol{z}_i - \boldsymbol{z}_j\|^2}$, and a repulsion force $(\boldsymbol{z}_i - \boldsymbol{z}_j)\frac{(1+\|\boldsymbol{z}_i - \boldsymbol{z}_j\|^2)^{-2}}{\sum_{r=1,r\neq i}^N(1+\|\boldsymbol{z}_r - \boldsymbol{z}_i\|^2)^{-1}}$. The interpretation of (10) follows the intuition of MAFE, that is at longer distances the embedding maps start to form clusters as determined by the attraction forces, while the repulsion forces are very negligible. As $\|\boldsymbol{z}_r - \boldsymbol{z}_i\| \to 0$ for each (i, r) pair, the repulsion magnitude dominates the interaction force vector. The convergence of the algorithm is established when the forces balance.

IV. New Superposed MAFE Model

A. Embedding Attractive Artificial Potential

The fundamental idea behind a superposed artificial field embedding is to treat the pair-equilibrium distances ϵ_{ij} for vertices in $\mathcal{G}_{\mathcal{S}}$ as attractive wells. Viewed another way, we think of the minimum-energy configuration between maps i and j as a *sink* of a potential function. Maps with high similarities in the observed neighborhood graph are pulled (by attraction forces) towards the common sink in the embedding space. The attractive potential functions that we consider can be seen as bounded from below to allow for the existence of constant attraction effects, that is $U_{att}^{ij}(\|\boldsymbol{z}_i - \boldsymbol{z}_j\|) \geq \alpha$, where α is a positive constant $\forall\ \|\boldsymbol{z}_i - \boldsymbol{z}_j\|$. For this paper, we choose attractive potentials of the form

$$U_{att}^{ij}(\|\boldsymbol{z}_i - \boldsymbol{z}_j\|) = \xi_a w_{ij}\|\boldsymbol{z}_i - \boldsymbol{z}_j\|^p \tag{10}$$

For values $0 < p \leq 1$, the pairwise attractive function is conic in shape and the resulting attractive force field has constant cluster formation amplitude determined from w_{ij} except at $\boldsymbol{z}_i = \boldsymbol{z}_j$, where it is singular. ξ_a is an attraction magnitude related parameter. Fig.2, shows the attractive potential energy generated from equation (10) for $p = 2, \xi_a = 1$, and $w_{ij} = 0.5$.

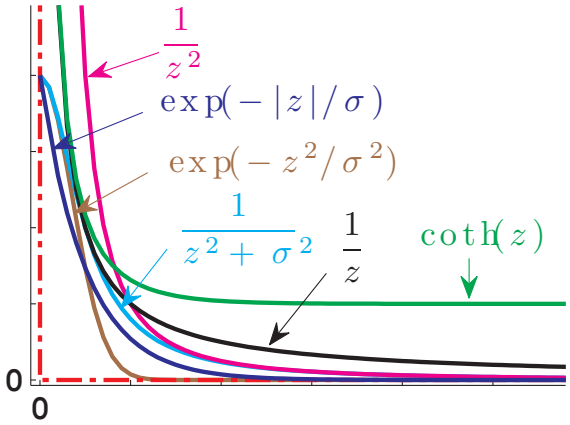


Fig. 1. Dashed lines show the function $I_+(\boldsymbol{z})$, and the solid curves show different forms of continuous decaying functions suitable for approximating $I_+(\boldsymbol{z})$.

B. Embedding Repulsive Artificial Potential

The basic idea in designing a repulsive potential function U_{rep}^{ij} is to think of an indicator function

$$I_+(\|\boldsymbol{z}_i - \boldsymbol{z}_j\|) = \begin{cases} 0 & \|\boldsymbol{z}_i - \boldsymbol{z}_j\| > \epsilon_{ij} \\ \infty & \|\boldsymbol{z}_i - \boldsymbol{z}_j\| \leq \epsilon_{ij}. \end{cases} \tag{11}$$

a nonincreasing function of distance. As the distance between pair-points increases, (11) is designed to have negligible influence on maps (*i.e.* maps are in long range zone where $F_r^{ij} < F_a^{ij}$). When the distance is small, the idea is to generate a barrier force between maps (*i.e.* maps are in the short range zone where $F_r^{ij} > F_a^{ij}$). Equation (11) best captures this notion. However it is not differentiable. We require its approximation by a differentiable function whose gradient can create a repulsion force F_r^{ij} with magnitude inversely proportional to the distance between pairs of maps *i.e.* $\|F_r^{ij}(\boldsymbol{z}_i - \boldsymbol{z}_j)\|_2 = \frac{1}{\mathbf{dist}(\boldsymbol{z}_i, \boldsymbol{z}_j)}$. Such approximations can be chosen from *e.g. Gaussian, Exponential, Cauchy, Hyperbolic Tangent* and *Inverse distance power* functions. The behavior of such functions in approximating $I_+(\|\boldsymbol{z}_i - \boldsymbol{z}_j\|)$ is shown in Fig.1.

1) Exponential Bounded Repulsion: The *Exponential or unnormalized Gaussian* curves in Fig.1 generates a continuous bounded approximation of (11). As the distance between maps grows, the magnitude of the curves approaches zero while a maximum magnitude is assigned for maps that get very close to each other. In this study, we propose a general bounded repulsion function of the form

$$U_{rep}(\|\boldsymbol{z}_i - \boldsymbol{z}_j\|) = \xi_r \sigma \exp\{-\frac{\|\boldsymbol{z}_i - \boldsymbol{z}_j\|^q}{\sigma}\} \tag{12}$$

where ξ_r is the repulsion magnitude related parameter. The Gaussian normalized version of this function appeared in (8). For $q = 2$, the function has spherical symmetry as shown in Fig.2. For values $0 < q \leq 1$, the repulsion potential field has the shape of a harmonic function often used in modeling obstacles in robotic path planning, while for $1 < q < 2$, it has the form of a tower centered at a point of interest.

C. MAFE-Bounded Repulsion Model

The superposed field with a bounded unnormalized Gaussian repulsion is given by

$$U(\mathcal{Z}) = \sum_{i,j \neq i} \xi_a w_{ij} \|\boldsymbol{z}_i - \boldsymbol{z}_j\|^p - \xi_r \sigma \exp\{-\frac{\|\boldsymbol{z}_i - \boldsymbol{z}_j\|^q}{\sigma}\}$$

Computing the gradient of yields the direction of motion for each individual pixel map position is described by

$$\begin{aligned} \dot{\boldsymbol{z}}_i &= -\sum_{j \neq i} (\boldsymbol{z}_i - \boldsymbol{z}_j) \xi_a w_{ij} p \|\boldsymbol{z}_i - \boldsymbol{z}_j\|^{p-2} \\ &- \xi_r q \|\boldsymbol{z}_i - \boldsymbol{z}_j\|^{q-2} \exp\{-\frac{\|\boldsymbol{z}_i - \boldsymbol{z}_j\|^q}{\sigma}\} \end{aligned}$$

An illustration of the gradient field for a point with strong attraction force field is shown in Fig.2. Without an attraction term, cluster formation would not occur since all maps would disperse from each other; whereas without repulsion, all maps would collapse to a single point leading to what's known as the crowding problem.

1) Embedding Space Weights: The search for minimum energy configuration establishes other additional properties of interest on the neighborhood graph, *i.e.* the lower dimensional space pairwise similarities $\tilde{w}_{ij}$. For the model designed with $p = q = 2$ in (13), the embedding neighborhood graph is described by

$$\tilde{w}_{ij} = \xi_a w_{ij} - \xi_r \exp\{-\frac{\|\boldsymbol{z}_i - \boldsymbol{z}_j\|^2}{\sigma}\} \tag{13}$$

In contrast to the high dimensional pairwise weights w_{ij}'s that are positive for all (i, j) in the observation graph, $\tilde{w}_{ij}$ can be negative as determined by the magnitude of the interaction force between the attractive and repulsive fields. This property establishes another point of departure by the general `MAFE` based techniques from traditional nonlinear embedding methods that enforce learning of positive lower dimensional weights *e.g.* `LLE, SNE, tSNE,` and `Isomap`.

V. Bilateral Kernel for Signature Similarities

Spatial preprocessing methods are often applied to remove noise and smooth images. These methods also enhance spatial texture information resulting in features that improve the performance of classification techniques. For example in [23], nonlinear diffusion partial differential equations (PDEs) and wavelet shrinkage were used for spatial preprocessing of hyperspectral images, and the results obtained demonstrated a significant improvement on classification performance. In this study, we adapt a bilateral filtering approach to devise a similarity function over the observed image pixels. The traditional formulation of a bilateral filter incorporates a linear convolution kernel, $K_s^{(i,j)} = \exp\left\{-\|\boldsymbol{s}_i - \boldsymbol{s}_j\|^2\right\}$, which weighs image pixel values as a function of the spatial distance from the center pixel , *and* also employs a nonlinear term $K_y^{(i,j)} = \exp\left\{-(\boldsymbol{y}_i - \boldsymbol{y}_j)^T \boldsymbol{\Sigma}_y^{-1} (\boldsymbol{y}_i - \boldsymbol{y}_j)\right\}$, which simply weighs pixel values as a function of the photometric differences between the center pixel and its neighbor pixels [25], [27], [28]. For illustration purposes, let us define the unnormalized similarity of pixels i and j as

$$w(\boldsymbol{s}_i, \boldsymbol{s}_j, \boldsymbol{y}_i, \boldsymbol{y}_j) = K_s^{(i,j)} \cdot K_y^{(i,j)} \tag{14}$$

where $\boldsymbol{s}_i$ denotes the spatial coordinates of pixel i, $\boldsymbol{y}_i$ denotes the photometric d-dimensional vector, with d corresponding to the number of spectral channels. Given N hyperspectral pixels, organized into a zero-mean data matrix $\boldsymbol{Y} = [\boldsymbol{y}_1 \boldsymbol{y}_2 \cdots \boldsymbol{y}_N] \in \mathbb{R}^{d \times N}$, the sample covariance is computed as $\boldsymbol{S} = \frac{1}{N}\boldsymbol{Y}\boldsymbol{Y}^T = \langle \boldsymbol{y}\boldsymbol{y}^T \rangle$, with the angle brackets denoting the average over N pixels. Thus, $\boldsymbol{S}$ is a $d \times d$ matrix whose diagonal components indicate the magnitude of noise variation in each of the d spectral channels, and the off-diagonal elements denote the extent to which noise co-vary with each pair of spectral bands. We make an observation that one can represent the unnormalized kernel $\mathcal{K}$ as a product of unnormalized gaussian functions, one for each pixel $\boldsymbol{y}_i$, yielding

$$\begin{aligned} \mathcal{K} &= \exp\left\{ \frac{-1}{2} \sum_{j=1}^{N} (\boldsymbol{y}_j - \boldsymbol{y}_i)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{y}_j - \boldsymbol{y}_i) \right\} \\ &= \exp\left\{ \frac{-tr(\boldsymbol{S}\boldsymbol{\Sigma}^{-1})}{2} + \sum_{j=1}^{N} \boldsymbol{y}_j \boldsymbol{\Sigma}^{-1} \boldsymbol{y}_i - \frac{N}{2} \boldsymbol{y}_i^T \boldsymbol{\Sigma}^{-1} \boldsymbol{y}_i \right\} \end{aligned}$$

where $\boldsymbol{\Sigma}^{-1}$ is the inverse covariance matrix, and $tr(\boldsymbol{B})$ denotes the trace of matrix $\boldsymbol{B}$. We could assume a zero-mean unnormalized Gaussian noise model over the pixels, i.e. we can simply subtract the center $\boldsymbol{y}_i$ from the data, to obtain a simplified expression as

$$\begin{aligned} \mathcal{K} &= \exp\left\{ \frac{-1}{2} \sum_{j=1}^{N} \boldsymbol{y}_j^T \boldsymbol{\Sigma}^{-1} \boldsymbol{y}_j \right\} \\ &= \exp\left\{ \frac{-1}{2} tr(\boldsymbol{Y}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{Y}) \right\} \end{aligned} \tag{15}$$

Note that $tr(\boldsymbol{Y}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{Y}) = tr(\boldsymbol{\Sigma}^{-1} \boldsymbol{Y}^T \boldsymbol{Y}) = N tr(\boldsymbol{\Sigma}^{-1} \boldsymbol{S})$. This shows that $\boldsymbol{S}$ is a sufficient statistic for characterizing the unnormalized likelihood (herein the photometric similarity) of data $\boldsymbol{Y}$, and we can further write

$$\mathcal{K} = \exp\left\{ \frac{-N}{2} tr(\boldsymbol{\Sigma}^{-1} \boldsymbol{S}) \right\} \tag{16}$$

We next consider a decomposition of the true covariance matrix into the product $\boldsymbol{\Sigma} = \boldsymbol{E}\Lambda\boldsymbol{E}^T$, where $\boldsymbol{E}$ is the orthogonal eigenvector matrix and $\boldsymbol{\Lambda}$ is the corresponding diagonal matrix of eigenvalues, to easily compute the covariance matrix whose inverse is required in (16). We adapt the efficient sparse matrix transform (SMT) approach in estimating the covariance matrix $\boldsymbol{\Sigma}$ [29]. The SMT approach solves the optimization problem, $\hat{\boldsymbol{E}} = \operatorname{argmin}_{\boldsymbol{E} \in \boldsymbol{\Omega}} \left\{ |\operatorname{diag}(\boldsymbol{E}^T \boldsymbol{S} \boldsymbol{E})| \right\}$, and set $\hat{\boldsymbol{\Lambda}} = \operatorname{diag}(\hat{\boldsymbol{E}}^T \boldsymbol{S} \hat{\boldsymbol{E}})$, where $\boldsymbol{\Omega}$ is the set of allowed orthogonal transforms that can be computed using a series of *Givens rotations* [29]. A simple manipulation can show that $\boldsymbol{\Sigma}^{-1} = \hat{\boldsymbol{E}} \hat{\boldsymbol{\Lambda}}^{-1} \hat{\boldsymbol{E}}^T$ so that we can perform unnormalized computations of the photometric distances and equate $K_y^{(i,j)} = \mathcal{K}$ in the transformed space. Thus,

$$K_y^{(i,j)} = \exp\left\{ -\frac{1}{2} (\hat{\boldsymbol{E}}^T \boldsymbol{y}_i - \hat{\boldsymbol{E}}^T \boldsymbol{y}_j)^T \hat{\boldsymbol{\Lambda}}^{-1} (\hat{\boldsymbol{E}}^T \boldsymbol{y}_i - \hat{\boldsymbol{E}}^T \boldsymbol{y}_j) \right\}$$

The SMT approach to computing the covariance matrix $\boldsymbol{\Sigma}$ is efficient and robust in handling the singularities of $\boldsymbol{\Sigma}$. Other approaches to computing $\boldsymbol{\Sigma}$ have been used in the literature including the PCA adaptation approach [25], where the singularity of $\boldsymbol{\Sigma}$ is not carefully addressed. The spectral signature similarities are stored in a $N \times N$ matrix $\boldsymbol{W} = [w(\boldsymbol{s}_i, \boldsymbol{s}_j, \boldsymbol{y}_i, \boldsymbol{y}_j)]$. Each row i of $\boldsymbol{W}$ is normalized by the sum of its elements to obtain a neighborhood probability distribution $W_i = [w_{ij}]$, which is the input to the embedding objective function.

VI. Optimization

Given the pair-dependent interactive *odd* functions, the objective function for MAFE(13) is simple to differentiate. For its optimization, we adapt a variation of the stochastic gradient descent [30] with common adaptive learning rate

$$\begin{aligned} \alpha^{(t+1)} &= \alpha^{(t)} + \gamma_1 \langle \nabla U(\mathcal{Z}^{(t-1)}), \nabla U(\mathcal{Z}^{(t)}) \rangle \\ &+ \gamma_2 \langle \nabla U(\mathcal{Z}^{(t-2)}), \nabla U(\mathcal{Z}^{(t-1)}) \rangle \end{aligned} \tag{17}$$

where $\alpha^{(t)}$ is the learning rate at iteration t, γ_1 and γ_2 are the meta-learning rates. The main characteristics of this fast learning rate adaptation scheme is that it exploits gradient-related information from the current as well as the two previous embedding coordinates in the sequence. This provides an enhancement on the stabilization in the values of the learning rate, and helps the gradient descent algorithm to exhibit fast convergence that leads to better minimum energy-configuration. A description of the proposed algorithm is given in Fig.1. The termination condition of the algorithm is when $\nabla U(\mathcal{Z}) \leq \epsilon$. The choices of γ_1 and γ_2 are not critical for finding the minimum-energy configuration, but only affect the rate at which we do so. Fig.3 shows smooth MAFE-BR gradient field trajectories during optimization as compared to both SNE and tSNE. In Fig.8, we present the convergence rates of this optimization scheme while contrasting with the rates obtained by the optimization methods used in SNE and tSNE.

Algorithm 1: MAFE Adaptive Stochastic Gradient Embedding

Input: Image data: $\boldsymbol{Y}$;
Initialize: $\alpha^{(1)}$, γ_1, γ_2;
Output: Embedding coordinates $\mathcal{Z} = \{\boldsymbol{z}_1^T, \boldsymbol{z}_2^T, \cdots, \boldsymbol{z}_N^T\}$;
Compute similarity weights $w_{ij} = K_s^{(i,j)} \cdot K_y^{(i,j)}$ (Eqn. (17));
$\mathcal{Z}^{(0)} \sim N(0, 50I)$;
Set $\mathcal{Z}^{(1)} = [\boldsymbol{z}_1^{(0)T}, \boldsymbol{z}_2^{(0)T}, \cdots, \boldsymbol{z}_N^{(0)T}]^T \in \mathbb{R}^{Nm}$;
while $\|\nabla U(\mathcal{Z}^{(t)})\| > \epsilon$ **do**
  Set $t = t + 1$;
  Compute new coordinates using;
  $\mathcal{Z}^{(t+1)} = \mathcal{Z}^{(t)} - \alpha^{(t)} \nabla U(\mathcal{Z}^t)$;
  Calculate the new learning rate from
  $\alpha^{(t+1)} = \alpha^{(t)} + \gamma_1 \langle \nabla U(\mathcal{Z}^{(t-1)}), \nabla U(\mathcal{Z}^{(t)}) \rangle +$
  $\gamma_2 \langle \nabla U(\mathcal{Z}^{(t-2)}), \nabla U(\mathcal{Z}^{(t-1)}) \rangle$;
end

2) Equilibrium State: We have so far been stating that the iterative optimization in (4), or in particular (13), will converge when all pair-equilibrium distances ϵ_{ij} are established. We can

make a strong theoretical argument that asserts that the motion of maps is guaranteed to stop and that no oscillatory behavior exists at the minimum-energy configuration state. This, we do by letting the invariant set of the equilibrium positions to be

$$\Xi_{equi} = \left\{ \mathcal{Z} : \dot{\mathcal{Z}} = 0 \right\}.$$

We can show that as $t \to \infty$, the state $\mathcal{Z}(t)$ converges to Ξ_{equi}, *i.e.* the minimum-energy configuration of the vertices position converges to a constant arrangement. This extends Theorem 2 of [21] to problems of dimension reduction and data visualization.

Theorem 1. *Consider a graph embedding described by* $\dot{\boldsymbol{z}}_{(i)} = \sum_{j \neq i} F^{ij}(\boldsymbol{z}_i - \boldsymbol{z}_j)$, $i = 1, \cdots, N$, *with force field function* $F^{ij}(\boldsymbol{z}_i - \boldsymbol{z}_j) = (\boldsymbol{z}_i - \boldsymbol{z}_j) \left\{ F_r^{ij}(\|\boldsymbol{z}_i - \boldsymbol{z}_j\|) - F_a^{ij}(\|\boldsymbol{z}_i - \boldsymbol{z}_j\|) \right\}$. *As* $t \to \infty$, *we have that* $\mathcal{Z}(t) \to \Xi_{equi}$.

Proof: Consider the general energy function $U(\mathcal{Z}) = \sum_{i=1}^{N} U_i(\mathcal{Z})$, where $U_i(\mathcal{Z})$ is defined in (2). Taking the derivative of $U(\mathcal{Z})$ with respect to each $\boldsymbol{z}_i$ yields

$$\begin{aligned} \nabla_{\boldsymbol{z}_i} U(\mathcal{Z}) &= \sum_{j \neq i} \left\{ \nabla_{\boldsymbol{z}_i} U_{att}^{ij}(\boldsymbol{z}_i, \boldsymbol{z}_j) - \nabla_{\boldsymbol{z}_i} U_{rep}^{ij}(\boldsymbol{z}_i, \boldsymbol{z}_j) \right\} \\ &= -\dot{\boldsymbol{z}}_i \end{aligned}$$

where we observe the negative gradient as direction of motion in the second equality. Taking the time derivative of $U(\mathcal{Z})$ along the motion of a graph configuration yields

$$\begin{aligned} \dot{U}(\mathcal{Z}) = \nabla_{\boldsymbol{z}_i} U(\mathcal{Z})^T \dot{\mathcal{Z}} &= 2 \sum_{i=1}^{N} \nabla_{\boldsymbol{z}_i} U(\mathcal{Z})^T \dot{\boldsymbol{z}}_i \\ &= 2 \sum_{i=1}^{N} \{ -\dot{\boldsymbol{z}}_i \}^T \dot{\boldsymbol{z}}_i \\ &= -2 \sum_{i=1}^{N} \|\dot{\boldsymbol{z}}_i\|^2 \leq 0, \quad \forall t. \end{aligned}$$

This result shows that the motion will continue in the direction of decreasing $U(\mathcal{Z})$ to a state when all $\dot{\boldsymbol{z}}_i = 0$. By invoking the Lasalle Invariance Principle [24], we can conclude that as $t \to \infty$ the graph configuration state $\mathcal{Z}(t)$ converges to the largest subset of the set defined as

$$\Xi = \left\{ \mathcal{Z} : \dot{U}(\mathcal{Z}) = 0 \right\} = \{ \mathcal{Z} : \dot{\boldsymbol{z}}_i = 0 \} = \Xi_{equi}.$$

Since each $\dot{\boldsymbol{z}}_i \in \Xi_{equi}$ is an equilibrium point, Ξ_{equi} is an invariant set and this concludes the proof. ∎

This general result holds for any function F^{ij} chosen based on the embedding force field properties discussed in Section II-A. It also extends to `tSNE` and `SNE` related formulations with a change of objective function. This result guarantees the convergence of the proposed algorithm. However, in practise, the termination condition is set to a finite time optimization of $U(\boldsymbol{z})$.

VII. Data Sets and Experiments

A. Botswana Hyperion

Hyperion data with 8 identified classes of complex natural vegetation were acquired over the Okavango Delta, Botswana, in May 2001, [13], [31]. The general class groupings include seasonal swamps, occasional swamps, and woodlands. Signatures of several classes are spectrally overlapped, typically resulting in poor classification accuracies. After removing water absorption, noisy, and overlapping spectral bands, 145 bands were used for classification experiments. We report on Euclidean and non-Euclidean embedding results as well as evaluation of the computed coordinates based on classification error rates for all 8 classes.

B. Kennedy Space Center (KSC)

Airborne hyperspectral data were acquired by the National Aeronautics and Space Administration(NASA) Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) sensor at 18-m spatial resolution over Kennedy Space Center during March 1996. Noisy and water absorption bands were removed, leaving 176 features for 13 wetland and upland classes of interest. Cabbage Palm Hammock (Class 3) and Broad Leaf/Oak Hammock (Class 6) are upland trees; Willow Swamp (Class 2), Hardwood Swamp (Class 7), Graminoid Marsh (Class 8) and Spartina Marsh (Class 9) are trees and grasses in wetlands. Their spectral signatures are mixed and often exhibit only subtle differences. The results for all 13 classes including these "difficult" classes are reported for the embedding and classification experiments.

C. Experiments

The experimental setup adapts the methodology used in the dimensionality reduction tool box [26] for the existing algorithms discussed. The setup consists of a principal component analysis (PCA) phase to reduce the dimension of the feature vectors to 40, from which the similarity values are computed using a Gaussian kernel function. The PCA dimension is experimentally set to suppress noise and speed up computation for the existing algorithms. In contrast, the proposed `MAFE-BR` algorithm achieves high quality results from a faster optimization scheme that does not include a PCA step. Additional comparison results are included to compare `MAFE-BR` and `MAFE-BR-PCA` to highlight the robustness of the proposed bilateral similarity kernel. Experiments are conducted using the model from equation (13) with a choice of $p = q = 2$, establishing a quadratic attraction and a spherical Gaussian repulsion model. Visualization results are obtained by mapping to a 2-dimensional space, with classification results generated for a varying lower dimensional feature space. Land cover label information for each pixel is only used to identify clusters hence the semi-supervised nature of the proposed embedding scheme. There is no consensus as to how should one evaluate a dimensionality reduction algorithm, a choice is often made depending on the application task. We further analyze the quality of the embedding representations by performing a 1NN classification measure on the data. A 1NN classifier is chosen simply because it ensues as an unbiased measure for comparing different embedding techniques. It is efficient, makes no assumption about class distribution and requires no model parameters to be set. Other sophisticated

TABLE I
EXPERIMENTAL DATA: CLASS LABELS AND NUMBER OF LABELED SAMPLES

Botswana Hyperion		Kennedy Space Center	
c1	Water (158)	c1	Scrub (761)
c2	Floodplain (228)	c2	Willow swamp (243)
c3	Riparian (237)	c3	Cabbage hamm (256)
c4	Firescar (178)	c4	Cabbage palm(252)
c5	Island interior (183)	c5	Slash pine (161)
c6	Woodlands (199)	c6	Oak (229)
c7	Savanna (162)	c7	Hardwood swamp (105)
c8	Short mopane (124)	c8	Graminoid marsh (431)
		c9	Spartina marsh (520)
		c10	Cattail marsh (404)
		c11	Salt marsh (419)
		c12	Mud flats (503)
		c13	Water (927)

methods including support vector machines (SVM) and maximum likelihood classifiers [18] could be used for classification even though they introduce additional bias in computing class boundaries which may not be a fair comparison of different embedding techniques. Additional evaluation results are also reported for the iterative gradient based techniques including MAFE−BR, where we compute the gradient field trajectories to examine the stability of the embedding map.

The optimization of MAFE−BR is terminated when $\|\nabla U(\boldsymbol{z}^{(t)})\| < \epsilon$, with $\epsilon = 10^{-5}$. The force field magnitude parameters are experimentally set between $0 < \xi_r, \xi_a < 1$ such that $\xi_a > \xi_r$ (*e.g.* $\xi_r = 10^-3$ and $\xi_a = 10^-2$) to maintain strong interaction force effects. For both tSNE and SNE, the gradient descent algorithm is run for $T = 1000$ iterations, while the value of σ is set following the approaches in [11], [12], whereby we first compute the perplexity of the conditional distribution induced by the Gaussian kernel determined as $2^{H(w)}$, where $H(w)$ is the entropy of the high dimensional neighborhood distribution. Embedding maps obtained by Isomap are based on the classical formulation that admits the closed-form solution of an eigen-structured problem, namely picking the leading components of variation. LE embedding solution is also based on solving an eigenvalue problem, but it relies on picking the trailing eigenvectors. In constructing the high dimensional neighborhood graph for Isomap and LE, we vary the number of neighbors from $k = 1$ to $k = 50$ and pick an optimal value of $k = 15$. This number ensures that the embeddings are neither too noisy and unstable nor does the geometry of the images exhibit a significant collapse of embedding coordinates.

1) Gradient Field Trajectories in Embedding Space: The motion dynamic equation in (1) serves as an approximation to the model that captures the formation of spectral signature manifolds. We obtain additional insights on the embedding algorithm by exploring the gradient vector fields as each map traverses towards the minimum configuration state of the graph. By plotting the changing vertex positions in 2-dimensional space, the trajectories reveal how the cluster formation gets affected by the optimization scheme. As an example, we consider the trajectories formed in mapping 15 hyperspectral pixels from three Botswana data classes. We randomly pick five samples from the (*Woodlands, Firescar,* and *Island Interior*) classes. Fig.3 shows MAFE−BR's optimization paths compared to the trajectories obtained for strategies used in tSNE and SNE. A close-up look on the trajectories in Fig.3(e) and Fig.3(g) shows that there is a high degree of colliding, poor learning rate, and instabilities associated with tSNE and SNE. This behavior is due to their objective cost functions that incorporate log sum terms in the repulsion potential term. Log sum term complicates the derivation of the gradient equations, and increases the degree of system nonlinearity. As a result, the simple optimization scheme used in tSNE and SNE takes long to establish the equilibrium state of the system. In contrast, the results in Fig.3(b) and Fig.3(c) show that MAFE−BR generates smooth trajectories with no sign of instability, i.e. no random change of gradient vector direction. The smooth force field interaction between the attraction and repulsion functions, and the optimization strategy lead to a stable balance which establishes the equilibrium state much quicker. MAFE−BR has several magnitudes of faster convergence speed as shown in Fig.8. In addition, it produces an embedding map where similar samples are close neighbors while setting clear boundaries between samples from different classes. Similar trends on gradient trajectory results were observed for the Kennedy Space Center data.

2) Visualization and Embedding Results: A more recent comparative study demonstrated that nonlinear dimension reduction methods often do better in capturing the structure of the associated manifolds in tasks with small number of classes than for problems with many classes [13]. This is mainly due to the complexity of the data manifolds represented in problems with many disparate classes. In many cases of hyperspectral data, similar classes of data may result in multiple manifolds due to their spatial location. This poses a challenge in many existing dimensionality reduction techniques that seek to map all similar or related data onto a single cluster, thereby increasing chances of collapsing different classes on top of each other. This is known as the *crowding problem* described in the introduction. In this section, we report on the performance of various embedding techniques in comparison to MAFE−BR combined with a bilateral similarity function, as well as MAFE−BR combined with a Gaussian kernel function.

The significant contrast between the results obtained when using these two similarity functions points out the importance of designing suitable kernel functions that capture geodesic relations along the data manifold. The proposed bilateral similarity function appears to be a strong suitable candidate for capturing such relations in hyperspectral images.

In Fig.4(a), we show the ground references of the Botswana data. Fig.5 shows the experimental results obtained with MAFE−BR, SNE, tSNE, Isomap, and LE on the Botswana data. MAFE−BR results display a radically different and superior embedding map compared to other methods. For example, SNE computes coordinates that seem to separate different classes well. However, there is a significant overlap on the *Riparian* and *Woodlands* classes, and the clusters seem to be more spread implying large variance. On the other hand, MAFE−BR has very tight spatial disjoint clusters, and no overlaps on the *Riparian* and *Woodlands* classes (except the seemingly touching boundaries for certain classes). These are the most difficult classes to separate in this data set as observed with the demonstration in [13]. MAFE−BR result is the strongest in capturing the ground truth map. tSNE has the capability to mitigate overcrowding of points, a good clustering effect. However, it leads to significant overlap between *Riparian* and *Woodlands* classes. In Isomap results we see significant overlaps for dissimilar classes. LE collapses the geometrical structure of the different features leading to poor separation of classes.

In Fig.4(b), we show the ground references of the Kennedy Space Center (KSC) data. Fig.7 shows the KSC embedding results obtained by all methods discussed. The embedding of this data set reveals a huge tendency of overcrowding and class overlapping by LE and Isomap. Almost all classes except for *Water* are not separable. tSNE and SNE representations shows significant levels of separation with the *Water, Salt Marsh,* and the *Spartina Marsh* classes. However, there is no visible distinction or separability characteristics for the remaining ten classes. In contrast, MAFE−BR constructs very compact and spatially-driven disjoint clusters capturing the land cover categories including their locations as displayed in the ground truth data. Furthermore, MAFE−BR has a *tiling* nature adaptability when classes are very close to each other. Fig.6 contrasts on the results obtained with MAFE−BR taking as input the neighborhood graph generated by the bilateral similarity function versus an approach which incorporates PCA and then computes the neighborhood graph for MAFE−BR−PCA. It is clear from the visual representation that PCA does influence the clusters to be well separated(increases the variance of maps). However, the within cluster compactness (tightness) is not preserved. The difference in the visual structures can be attributed to the additional information that is neglected by PCA as it retains the details corresponding to only the largest 40 eigenvalues. In addition, one would hope that a PCA step will enable a faster computation for the embedding algorithm since the level of noise is reduced. However, Fig.8 shows our algorithmic runtime evaluation suggesting that there is no benefit in computational speed when a MAFE−BR model incorporates PCA over use of all dimensions in the samples for generating the neighborhood graph.

3) Classification Results: Classification results are included as further evaluation of the representations obtained by different embedding techniques in comparison to the proposed MAFE−BR approach. Embedded pixel maps were randomly sampled to generate 60% training and 40% testing samples, with results averaged over 10 runs. All approaches were compared on the same samples to maintain a consistent comparison. We have also added two more approaches: the largest margin nearest neighbor (LMNN) [14], which is primarily not an unsupervised dimensionality reduction method but a classification technique based on the notion of using training class labels of largest margin neighbors to compute a metric for inferring the decision boundaries during testing. We also included the local Fisher discriminant analysis(LFDA) [4], a dimensionality reduction technique whose embedding solution is based on solving an eigen-structure problem. During the experimental exercise, we noted that the embedding solutions obtained by LFDA and MDS were similar to the Isomap solution, hence their exclusion from the presentation.

Tables II and III illustrate the 1NN semi-supervised classification performance accuracy per class. The trends observed with a bilateral kernel(BK) neighborhood graph shows a clear separation of classes, and as a result the representations achieves an increased classification accuracy as demonstrated with MAFE-BR-BK. In contrast, classification results obtained with a MAFE-BR with a Gaussian kernel neighborhood graph demonstrates a similar performance to SNE. By ignoring all terms that do not depend on the embedding positions, SNE differs from MAFE-BR only in the scaling of the repulsion potential energy. In MAFE-BR, the potential energy function is less nonlinear, and this leads to simple gradient equations that are easier to derive in contrast to SNE derivations [12]. The difference between MAFE-BR-BK and MAFE-BR classification accuracy highlights the potential benefit that the proposed bilateral kernel has on other algorithms including SNE and tSNE albeit their complicated derivations.

The Botswana data classification has the following insights that are consistent with the visualization maps from the previous section: all methods seem to achieve a better per class performance accuracy, while the lowest accuracy results are achieved with the LE embedding representation. Furthermore, we have the lowest accuracy per class between class 3 (c3) and class 6 (c6) corresponding to the *Riparian* and *Woodlands* classes, respectively. Lower classification results on c3 and c6 are expected because these two classes are the most difficult to separate, in consistency with the results shown in Fig.5(similarly as demonstrated in [13], [33], [34]). LMNN using 1NN achieves the second best results owing to its ability to make use of class label information in learning the Mahalanobis distance metric for 1NN classification. The objective function for LMNN does have a force field structure in which class label information is used to compute the optimal metric with a goal that k-nearest neighbors always belong to the same class (*i.e*. pulled closer by an attraction term) while example samples from other classes are separated by a large margin (*i.e*. pushed far by a repulsion term). In contrast MAFE−BR's objective function is formulated as a function of the distance between pairs of points, and no class label

information is used during computation of the maps. For the KSC data classification, a similar trend is observed with MAFE-BR-BK providing a coordinate representation from which a higher 1NN classification performance is achieved. The classification results achieved for class 3(c3), class 4 (c4), class 5(c5), and class 6(c6) indicate the lowest performance in all embedding spaces except for the solution achieved by MAFE-BR. From the visualization result shown in Fig.7, these classes correspond to the *Cabbage Palm Hammock, Cabbage Palm/Oak Hammock, Slash Pine,* and *Oak/Broadleaf Hammock*, respectively. These are all categories of very similar upland trees. Their spectral signatures are mixed, and often exhibit only subtle differences. However, a combination of MAFE-BR and a spatially-sensitive bilateral similarity function shows that even with complex mixed classes, the proposed graph embedding algorithm does separate difficult land cover categories with a high degree of accuracy which is reflected in the classification results.

In Fig.9, we show the mean $\pm$ one standard error misclassification error plots as a function of the embedding dimension, *i.e.* $m = 1 \sim 20$. The question of how to choose the optimal dimension of the embedding space is addressed by adapting a manifold projection approach. A manifold projection is based on first defining the spatial-spectral neighborhood graph structure of the data. If each neighborhood can be projected to an m-dimensional space within a classification tolerance, then the intrinsic dimension of the data is m, and we can reduce the dimension to m without losing much information. For Isomap, MDS, LE, tSNE, and SNE the manifold projection result is based on experimentally setting the number of neighbors in the graph. If the size of the neighborhood is too small, then the geometry of the data will cause important data features to be collapsed e.g. the result of LE in Fig.5 and Fig.7. On the other hand, if the neighborhood is too large, which creates a large intrinsic dimension, then the manifold projections could become noisy and unstable. In contrast, MAFE-BR-BK mitigates such challenges by automatically introducing a spatially induced sparsity structure on the neighborhood graph with the number of neighbors selected depending on the disjoint nature of the data. Using Fig.9(b) and Fig.9(a), we can estimate the optimal dimension for mapping of both the 145 dimension Botswana spectral channels and the 176 Kennedy Space Center spectral channels as $m = 8$. There are other methods in the literature for estimating the embedding dimension. For example, the virtual dimensionality method which defines the minimum number of spectrally distinct signal sources that characterize the hyperspectral data is introduced in [16], and an estimation based on subspace identification can be found in [17]. Some very promising approaches include the local image background detection for signature-based object detection applications [10] and the projection approach based on local information for subspace-based detection of spectral anomalies [9].

TABLE II
CLASSIFICATION RESULTS CARRIED OUT IN DIFFERENT EMBEDDING SPACES FOR BOTSWANA DATA.

Class	Accuracy on single class (%)								
	MDS	SNE	tSNE	MAFE-BK**	MAFE-GK	Isomap	LE	LMNN**	LFDA
c1	100	100	100	**100**	100	100	100	**100**	100
c2	98.74	98.74	96.86	**100**	99.17	100	97.06	**100**	100
c3*	**89.7**	**89.09**	**88.48**	**96**	**88.83**	**91.55**	**80.28**	**94.37**	**91.55**
c4	97.6	97.6	97.6	**100**	97.51	100	96.30	**100**	98.11
c5	96.09	94.53	95.31	**100**	93.57	96.36	80	**100**	94.55
c6*	**92.86**	**92.14**	**90**	**100**	**91.80**	**89.83**	**69.49**	**98.31**	**94.92**
c7	96.49	97.37	96.49	**100**	97.14	97.92	77.08	**97.92**	97.92
c8	96.55	96.55	96.55	**100**	96.6	100	94.59	**100**	97.37
KS	95.08	94.75	93.97	**99.33**	94.68	95.89	88.45	**98.82**	95.55

* The most difficult classes to separate.
** Embedding space(s) providing highest 1NN classification accuracy.

VIII. DISCUSSION AND FUTURE WORK

The experimental results presented demonstrate that the disjoint characteristics inherent in hyperspectral data can be

TABLE III
CLASSIFICATION RESULTS CARRIED OUT IN DIFFERENT EMBEDDING SPACES FOR KSC DATA.

Class	Accuracy on single class (%)								
	MDS	tSNE	MAFE-BK**	MAFE-GK	SNE	Isomap	LE	LMNN	LFDA
c1	90.91	90.34	**97.16**	90.71	91.2	9.33	72	93.33	92
c2	90.74	87.04	**100**	86.4	85.34	62.22	43.48	86.96	82.61
c3*	**80**	**78.33**	**100**	**81.7**	**80.4**	**80.77**	**38.48**	**76.92**	**88.46**
c4*	**57.89**	**45.61**	**84.21**	**54.3**	**51.9**	**56**	**28**	**68**	**68**
c5*	**51.28**	**41.03**	**100**	**41.3**	**41.20**	**47.06**	**29.41**	**41.18**	**52.94**
c6*	**39.29**	**35.71**	**82.14**	**39.10**	**39.30**	**40**	**28**	**84**	**72**
c7	80.77	80.77	**100**	80.77	82.77	90.91	36.36	90.91	100
c8	74.22	60	**100**	66.2	63.62	69.05	35.71	90.48	88.10
c9	94.21	91.74	**100**	93.2	93.56	96.15	75	90.38	96.15
c10	89.47	93.68	**100**	94.6	93.68	92.68	80.49	100	100
c11	93.81	93.81	**98.97**	95.6	93.81	97.56	85.37	95.12	97.56
c12	82.20	80.51	**100**	80.9	81.91	92.16	80.39	98.04	98.04
c13	100	100	**100**	100	100	100	98.39	100	100
KS	83.04	80.10	**97.86**	82.10	81.90	86.40	76.71	92.1	89.2

* The most difficult classes to separate.
** Embedding space(s) providing highest 1NN classification accuracy.

mapped to lower dimensional spaces by first characterizing the spectral signature relations using a sparse matrix transformed spatially-sensitive neighborhood graph. The graph weights are derived from a joint spatial and photometric distance based bilateral kernel function that improves the capability to capture regularities and enhances the similarities within high dimensional spectral signatures. The bilateral similarity function is general in its form, and could potentially be used to benefit existing algorithms including the ones discussed in this study. Herein the neighborhood relations are mapped onto a low dimensional space based on the notion of a force field graph embedding formulation. Graph embedding is performed under a framework that promotes representing local relations with small distances while global relations are modeled by longer distances. Meaningful structures emerge under the general framework as a result of the interplay between the attraction and the repulsion forces that exist for every pair of vertices on the neighborhood graph. As an alternative framework for dimensionality reduction and visualization of data, MAFE encodes the intrinsic geometries underpinning the nonlinear characteristics in the data, and achieves better representation as evaluated on tasks that included visualization and classification of various commonly studied hyperspectral imagery.

Dimensionality reduction and visualization of data often requires a trade-off between accuracy and computational efficiency. As demonstrated, encoding of similarity relations onto a neighborhood graph does have significant outcomes on the quality of the embeddings. Estimating the parameters of the kernel function often is a cumbersome task and may lead to increased computational hurdles. Even though our approach does achieve quality visualizations and increased classification accuracy over existing approaches, we still face the equal and similar computational and memory demands that are $O(N^2)$, where N is the number of observations. Such algorithmic inefficiencies are common in most techniques that do not provide closed form solutions, *e.g.*tSNE, SNE, Iterative-MDS. On the other hand, spectral based algorithms (*e.g.*LE, LLE, MDS) do conquer this shortcoming by providing closed form solutions based on eigendecomposition techniques. However, the accuracy of spectral based embeddings are very poor, as demonstrated in our presentation.

MAFE-BR may have smooth trajectories and a faster convergence, but it also introduces what we observed as *dynamic local maxima behavior* during cluster formation. The formation of clusters creates local repulsions leading to local traps for pixel maps that still need to move closer to their closest neighbors. As such, a weak choice on the attraction potential may lead to very poor distance preserving embeddings and may affect the convergence of the optimization algorithm. We are currently studying local maxima generated traps using piecewise potential functions that vary at different distances to increase attraction magnitudes in hope of overcoming dynamic local repulsion forces. Additional avenues worth investigating include developing theoretical analysis further, an often missing component of nonlinear dimension reduction algorithms. Further work could exploit the sparsity structure of the neighborhood graph to prescribe an efficient optimization framework.

IX. Conclusions

The main goal of the study was to develop a nonlinear embedding framework that preserves the neighborhood relations of highly nonlinear and disjoint structures. As an example, the disjoint characteristics inherent in hyperspectral data were mapped to lower dimensional space by first characterizing the spectral signature relations using a sparse matrix transformed spatially-sensitive neighborhood graph. The graph weights are derived from a joint spatial and photometric distance based bilateral kernel function that improves the capability to capture regularities and enhances the similarities within high dimensional spectral signatures. Adapting a force field intuition from mechanics, a dynamic system was derived. In this framework, pairwise interactions of moving particles (or maps) were assumed to determine both their positions (embedding coordinates) and the description of the lower dimensional space (neighborhood graph weights). We showed that the new embedding technique has often sought after desirable properties in preserving the local topology of spectral channels and also reveals natural global structures, *i.e*. disjoint clusters for hyperspectral imagery. The framework yields formulations of well known state-of-the-art dimensionality reduction techniques with very few assumptions, and could potential be used to derive new embedding models. Experimental work conducted on visualization, gradient field trajectories and classification of images acquired by multiple sensors at various spatial resolutions over different types of land covers indicates that a `MAFE-BR-BK` embedding representation outperforms other techniques.

Acknowledgment

The authors thank Dr M. Crawford and Lexie Yang for fruitful discussions and suggestions on the datasets. The knowledge and feedback comments from anonymous reviewers and editors were instrumental in compiling the final manuscript. This research was supported by the CSIR Meraka Institute of South Africa.

References

[1] C. M. Bachmann, T. L. Ainsworth,and R. A. Fusina, Exploiting manifold geometry in hyperspectral imagery, *IEEE Transactions on Geoscience and Remote Sensing*, vol.43, no.3, pp.441-454, 2005.

[2] I. Joliffe, Principal Component Analysis, *Springer-Verlag*, 1986.

[3] W. Torgerson, Multidimensional Scaling I: Theory and method, *Psychometrika*, vol.17,no.4,pp.401-419, 1952.

[4] M. Sugiyama, Dimensionality reduction of multimodal labeled data by local Fisher discriminant analysis, *Journal of Machine Learning Research (JMLR)*,Vol.8, pp.1027-1061, 2007.

[5] M. Sugiyama and T. Idé and S. Nakajima and J. Sese, Semi-Supervised Local Fisher Discriminant Analysis for Dimensionality Reduction, *Machine Learning Research*,Vol.78,no.1-2, pp.35-61, 2010.

[6] L. Song and A.J. Smola and K. Borgwardt and A. Gretton, Colored Maxium Variance Unfolding, *Advances in Neural Information Processing Systems*,Vol.20, pp.1385–1392, 2008.

[7] S. Roweis and L. K. Saul, Nonlinear dimensionality reduction by locally linear embedding, *Science*, vol. 290, pp.2323-2326, 2000.

[8] J. B. Tenenbaum, V. de Silva and J.C. Langford, A global geometric framework for nonlinear dimensionality reduction, *Science*, vol.290, pp.2319-2323, 2000.

[9] M. Hasanlou, and F. Samadzadegan, Comparative Study of Intrinsic Dimensionality Estimation and Dimension Reduction Techniques on Hyperspectral Images Using K-NN Classifier, *IEEE Geoscience and Remote Sensing Letters*, vol.9, no.6, 2012.

[10] S. Matteoli, N. Acito, M. Diani, and G. Corsini, An Automatic Approach to Adaptive Local Background Estimation and Suppression in Hyperspectral Target Detection, *IEEE Trans. on Geoscience and Remote Sensing*, vol.49, no.2, 2011.

[11] L. van der Maaten and G. Hinton, Visualizing data using t-sne, *JMLR*, vol.9, pp.2579-2605, 2008.

[12] G. Hinton and S. Roweis, Stochastic neighbor embedding, *ICML*, pp.833-840, 2002.

[13] M. M. Crawford, W. Kim, and M. Li, Exploring Nonlinear Manifold Learning for Classification of Hyperspectral Data, *Transactions on Optical Remote Sensing* , vol.3, pp.207-234, 2011.

[14] K.Q. Weinberger and L. K. Saul, Distance Metric Learning for Large Margin Nearest Neighbor Classification, *JMLR*,Vol.10, pp.207-244, 2009.

[15] M. Belkin, and P. Niyogi, Laplacian eigenmaps for dimensionality reduction and data representation, *Neural Comp.*,Vol.15, no.6,pp.1373–1396, 2003.

[16] C.-I Chang and Q. Du, Estimation of number of spectrally distinct signal sources in hyperspectral imagery, *IEEE Trans. on Geosci. and Remote Sens.*, vol. 42, no. 3, pp. 608-619, 2004.

[17] J. M. Bioucas-Dias and J. M. P. Nascimento, Hyperspectral subspace identification, *IEEE Trans. on Geosci. and Remote Sens.*, vol. 46, no. 8, pp. 2435-2445, 2008.

[18] L. Wei, S. Parasad, J. E. Fowler, and L. M. Bruce, Locality-Preserving Dimensionality reduction and Classification for Hyperspectral Image Analysis, *IEEE Trans. on Geosci. and Remote Sens.*, vol. 50, no. 4, pp. 1185-1198, 2012.

[19] D. Lunga and O. Ersoy, Spherical stochastic neighbor embedding of hyperspectral data, *IEEE Transactions on Geoscience and Remote Sensing*, (*in press*), 2012.

[20] V. Gazi, and K. M. Passino, Stability analysis of swarms, *Proc. American Control Conf.*, Anchorage, Alaska, May, 2002.

[21] V. Gazi, and K. M. Passino, Stability analysis of swarms, *IEEE Transactions on Automatic Control*, vol.48, no.4, pp. 692-697, April, 2003.

[22] J.C. Latombe, *Robotic Motion Planning*, Kluwer Academic Publishers Boston, MA, 1991.

[23] S. Velasco-Forero and V. Manian, Improving Hyperspectral Image Classification Using Spatial Preprocessing, *IEEE Transactions on Geoscience and Remote Sensing*, vol.6, no.2, pp. 297-301, 2009.

[24] J.P. LaSalle, Some extensions of Liapunov's second method, *IRE Transactions on Circuit Theory*, CT-7, pp. 520527, 1960

[25] P. Honghong and R. Raghuveer, Hyperspectral image enhancement with vector bilateral filtering, *IEEE international conference on Image processing*, pp.3669-3672, 2009.

[26] L.J.P. van der Maaten, E.O. Postma, and H.J. van den Herik., Dimensionality Reduction: A Comparative Review, *Tilburg University Technical Report, TiCC-TR 2009-005*, 2009.

[27] K. Kotwal and S. Chaudhuri, Visualization of Hyperspectral Images Using Bilateral Filtering, *IEEE Transactions on Geoscience and Remote Sensing*, vol.48, no.5, pp.2308-2316 ,May, 2010.

[28] C. Tomasi and R. Manduchi, Bilateral filtering for gray and color images, *IEEE International Conference on Computer Vision*, pp.839-846, 1998.

[29] J. Theiler, G. Cao, L. Bachega and C. Bouman, Sparse Matrix Transform for Hyperspectral Image Processing, *IEEE Journal of Selected Topics in Signal Processing*, vol.5, no.3,pp.424-437, June, 2011.

[30] Plagianakos, V P and Magoulas, G D and Vrahatis, M N' Chapter 2 Learning rate adaptation in stochastic gradient descent , Information Systems Journal, Kluwer Academic Pub, pp.15-26, 2001.

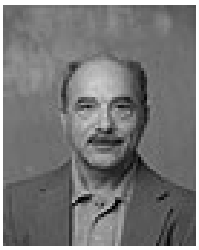
[31] A. L. Neuenschwander, Remote sensing of vegetation dynamics in response to flooding and fire in the Okavango Delta-Botswana, *Ph.D. dissertation, Univ. Texas Austin, Austin, TX*, 2007.

[32] J. Cohen, A coefficient of agreement from nominal scales, *Education Psychological Measurement* ,vol.20, no.1, pp.37-46, 1960.

[33] W. Di, and M. M. Crawford, Active Learning via Multi-View and Local Proximity Co-Regularization for Hyperspectral Image Classification, *IEEE Journal of Selected Topics in Signal Processing*, vol.5, no.3, pp.618-628, June, 2011.

[34] L. Ma, M. M. Crawford, and J. Tian, Local Manifold Learning-Based k-Nearest-Neighbor for Hyperspectral Image Classification, *IEEE Transactions on Geoscience and Remote Sensing*, vol.48, no.11, pp.4099-4109, 2010.

Dalton Lunga (S'08) received his B.Eng. and M.S. degrees in electrical engineering from the University of Johannesburg and the University of Witwatersrand, Johannesburg, South Africa, in 2004 and 2006, respectively. He earned an M.S.E.E. in 2011 and is currently pursuing a PhD in Electrical and Computer Engineering, at Purdue University, West Lafayette. His research interests include statistical signal processing, machine learning, optimization, manifold learning, remote sensing, image reconstruction and segmentation.



Okan Ersoy (M'86-SM'90-F'00) received the B.S.E.E. degree from Robert College, Istanbul, Turkey, in 1967, and the M.S. Certificate of Engineering, M.S., and Ph.D. degrees from the University of California, Los Angeles, in 1968, 1971, and 1972, respectively. He is currently a Professor in the School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN. His current research interests include remote sensing, machine learning and pattern recognition, digital signal/image processing and recognition, transform and time-frequency methods, imaging, diffractive optics, and distant learning. He has published more than 250 papers in his areas of research. He is also the holder of five patents. Dr. Ersoy is a Fellow of the Optical Society of America, and a fellow of IEEE.

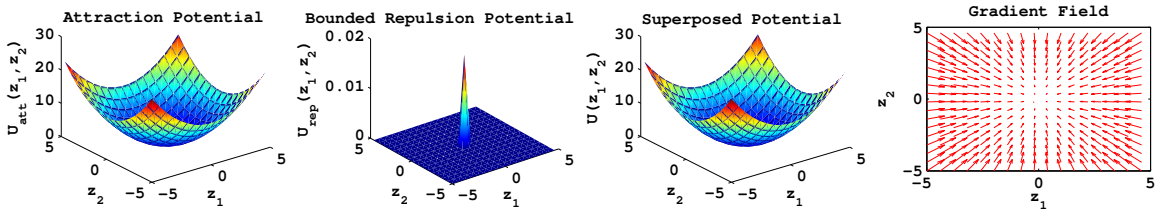
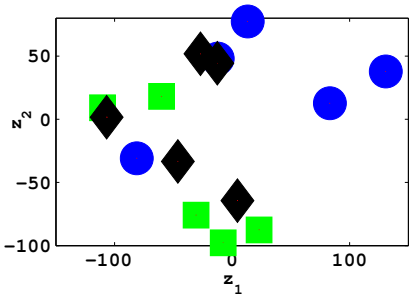
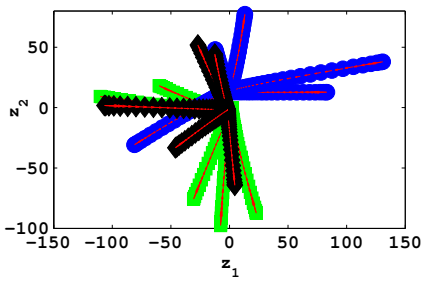


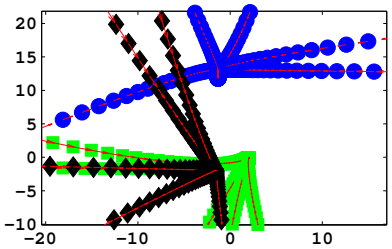
Fig. 2. Illustrations of a superposed-potential function for $p = q = 2$. Arrows indicate the negative gradient force field. The map located at $(z_1, z_2) = (0, 0)$ has a very strong attraction force over a large range of distance. The repulsion force is significantly small and over a very short range for this illustration.



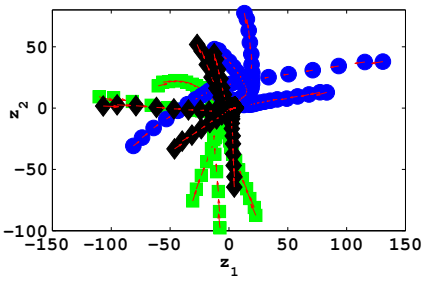
(a) Initial Maps.



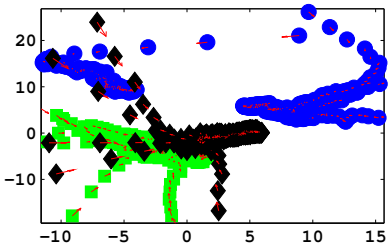
(b) MAFE-BR.



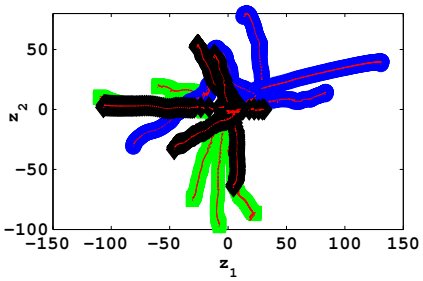
(c) MAFE-BR zoom-in.



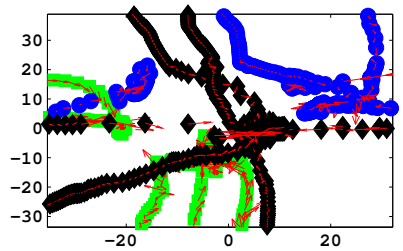
(d) SNE.



(e) SNE zoom-in.

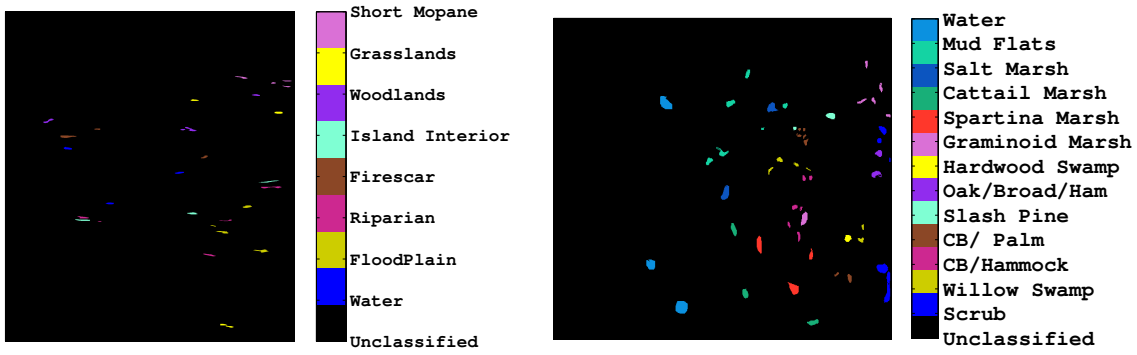


(f) tSNE.



(g) tSNE zoom-in.

Fig. 3. Gradient based optimization illustrations over 100 iterations. MAFE-BR(b) and (c), SNE(d) and (e), tSNE(f) and (g), are all initialized from the same seed of 15 points for Botswana maps of 3 different classes (*Woodlands, Firescar* and *Island Interior*). MAFE-BR displays smooth gradient trajectories. Similar points cluster and traverse in the direction of the gradient field as indicated by the arrows. SNE and tSNE trajectories are subject to oscillations, including collisions of maps leading to instabilities as well as the severe local maxima traps that slow down the optimization algorithm. Arrows are shown pointing in the negative direction of the gradient.



(a) (b)

Fig. 4. Class sample location for -(a) Botswana data and (b) Kennedy Space Center data.

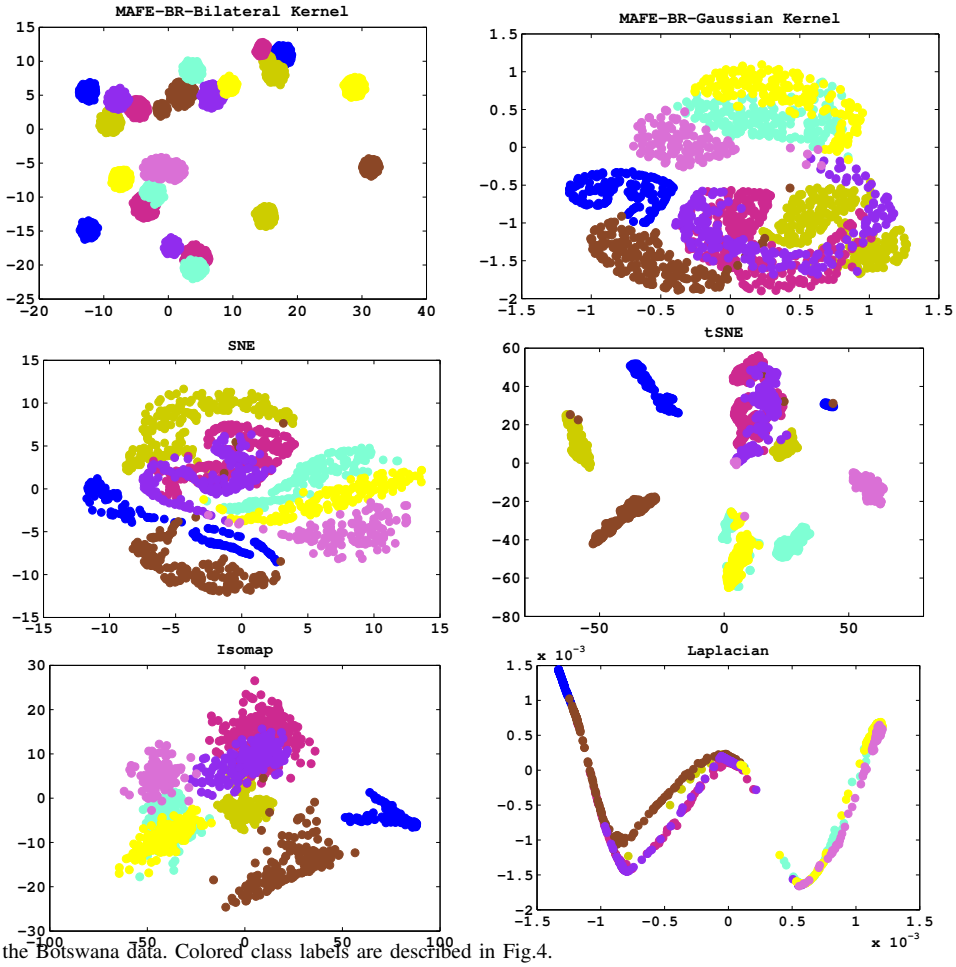


Fig. 5. Embedding of the Botswana data. Colored class labels are described in Fig.4.

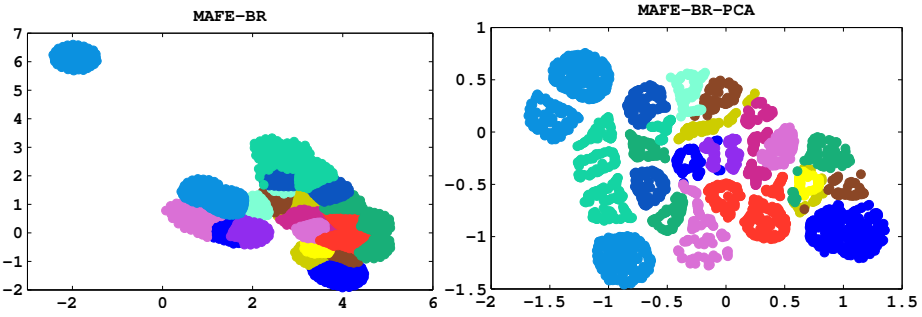


Fig. 6. MAFE-BR and MAFE-BR-PCA Embedding of Kennedy Space Center data.

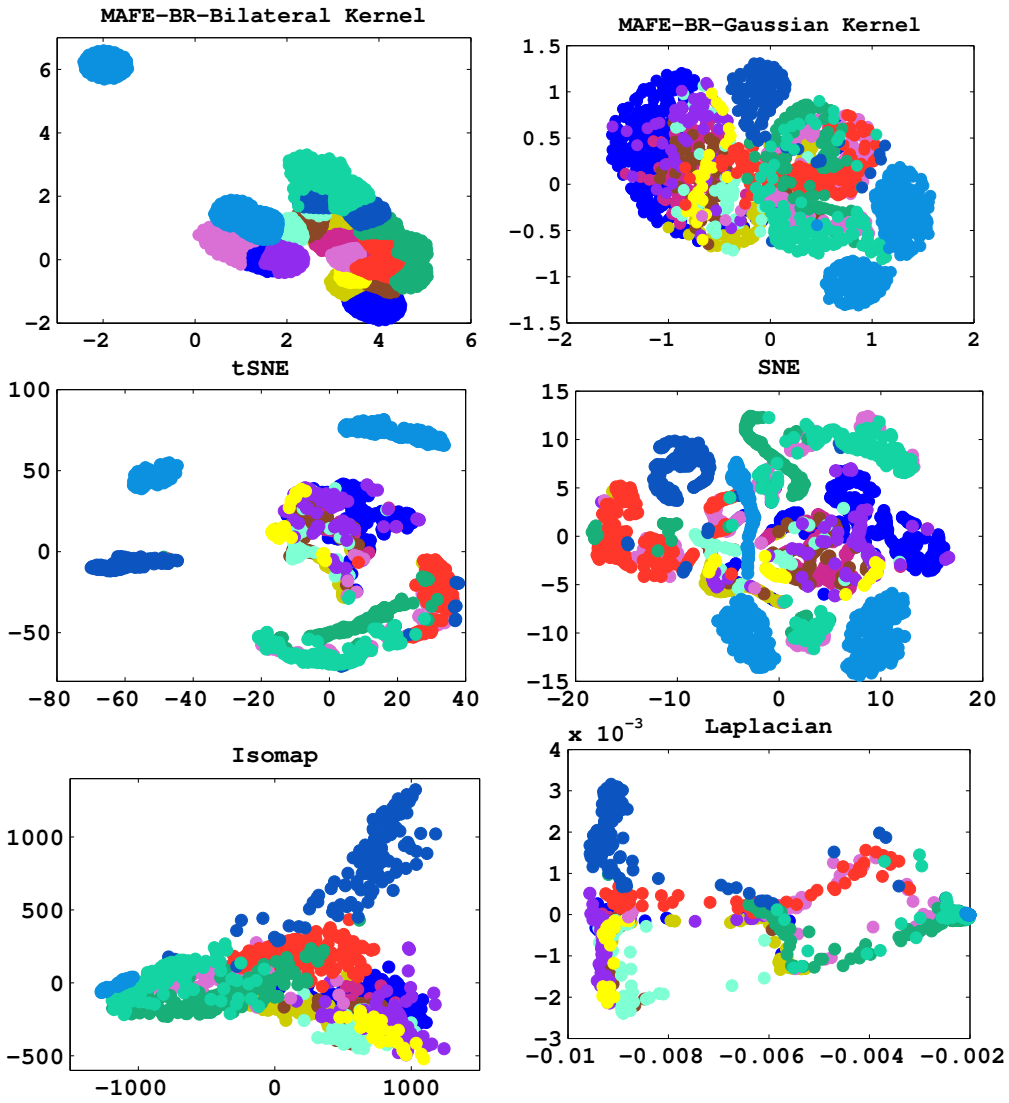
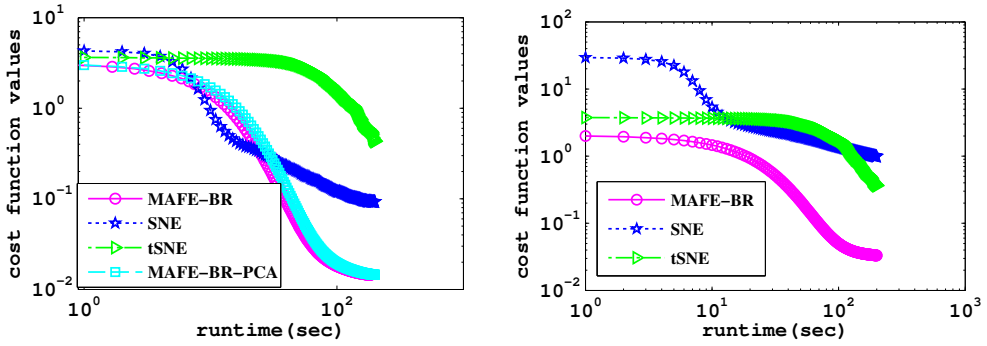


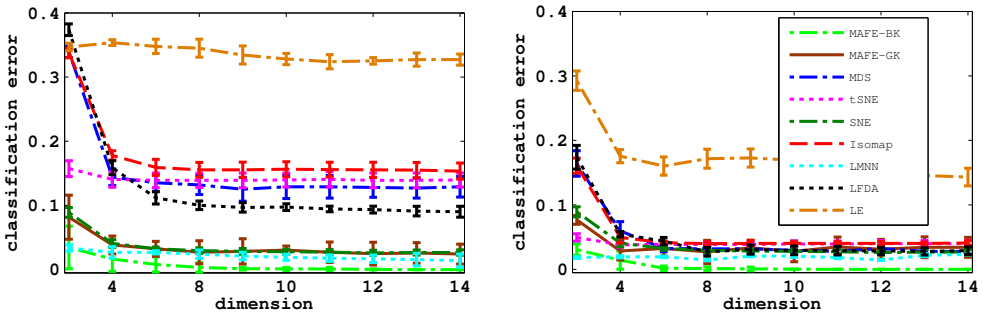
Fig. 7. Embedding of Kennedy Space Center data. Colored class labels are described in Fig.4.



(a) Embedding 3126 samples of KSC data

(b) Embedding 1469 samples of Botswana data

Fig. 8. Optimization run times for MAFE, SNE and tSNE algorithms.



(a) (b)

Fig. 9. Mean $\pm$ one standard error misclassification error comparison for 1-nearest neighbor classifier based on various embedding spaces while varying the dimension. (a) Kennedy Space Center data and (b) Botswana data.