MANIFOLD LEARNING BASED FEATURE EXTRACTION FOR CLASSIFICATION OF HYPERSPECTRAL DATA

D. Lunga[1], Member, IEEE. S. Prasad[2], Member, IEEE, M. Crawford[3], Fellow, IEEE, O. Ersoy[3], Fellow, IEEE

1. Meraka Institute, Council for Scientific and Industrial Research, South Africa.
2. Department of Electrical and Computer Engineering, University of Houston.
3. Schools of Civil Engineering and Electrical and Computer Engineering, Purdue University.

Interest in manifold learning for representing the topology of large, high dimensional nonlinear data sets in lower, but still meaningful dimensions for visualization and classification has grown rapidly over the past decade, and particularly in analysis of hyperspectral imagery. High spectral resolution and the typically continuous bands of hyperspectral image (HSI) data enable discrimination between spectrally similar targets of interest, provide capability to estimate within pixel abundances of constituents, and allow direct exploitation of absorption features in predictive models. Although hyperspectral data are typically modeled assuming that the data originate from linear stochastic processes, nonlinearities are often exhibited in the data due to the effects of multipath scattering, variations in sun-canopy-sensor geometry, nonhomogeneous composition of pixels, and attenuating properties of media. [1]. Because of the dense spectral sampling of HSI data, the associated spectral information in many adjacent bands is highly correlated, resulting in much lower intrinsic dimensions spanned by the data (Fig. 1). Increased availability of HSI and greater access to advanced computing have motivated development of specialized methods for exploitation of nonlinear characteristics of these data. In this context, feature selection and feature extraction approaches for dimensionality reduction have received significant attention. While both feature selection and extraction result in some loss of information relative to the original data, both have been demonstrated to be quite successful in the classification arena. While feature selection retains meaningful features for classification, the algorithms are computationally intensive and are often not robust in complex scenes. Alternatively, feature extraction approaches, which project the data to lower dimensional intrinsic spaces, are typically more robust to variation in spectral signatures across scenes, and most are computationally superior to optimal feature selection, although the interpretation relative to the original spectral signatures is lost. Both feature selection and extraction are flexible relative to the choice of the back-end classifier.

Theoretical contributions and applications of manifold learning have progressed in tandem, with new results providing capability for data analysis, and applications highlighting limitations in existing methods. For HSI, the enormous size of the data sets and spatial clustering of classes on the image grid provide both challenges and opportunities to extend traditional manifold learning methods. The machine learning community has demonstrated the potential of manifold based approaches for nonlinear dimensionality reduction and modeling of nonlinear structure [2]–[10]. The potential value of manifold learning for HSI analysis has been demonstrated for applications including feature extraction [1], [11], segmentation [12], classification [13]–[15], anomaly detection [16], [17], and spectral unmixing [18]–[21] with some approaches exploiting inter-band correlation [14], [15] and local spatial homogeneity [21]. Challenges encountered in analyzing data sets, have inspired recent advances in manifold learning methods, particularly related to feature extraction and visualization. This paper provides both an overview of traditional approaches and new directions for modeling HSI data on nonlinear manifolds.

A general framework for representing spectral signatures based on graph weights is presented, and traditional unsupervised global and local graph based methods for dimensionality reduction are summarized. Extensions to exploit labeled data in single image and multi-temporal sequences of hyperspectral data are described. Variants of manifold learning based projection are particularly suitable as a preprocessing to traditional Bayesian classification. In this context, locality preserving discriminant analysis methods are discussed. While traditional eigen-decomposition based methods are computationally advantageous, iterative methods can often provide improved separation of classes in the embedded space for both visualization and classification. Iterative methods are introduced in the context of the affinity matrix, which is utilized to describe Multidimensional Artificial Field Embedding (MAFE) and Spherical Stochastic Neighbor Embedding (SSNE) [22]. Examples of selected methods applied to a testbed of hyperspectral data are included for illustration of the methods using a 1-nearest neighbor classifier.

NONLINEAR MANIFOLD LEARNING IN A GRAPH EMBEDDING FRAMEWORK

Given a dataset with training samples $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n$ in $\mathbb{R}^m$ (m-dimensional feature space) and n is the total number of training samples, nonlinear dimensionality reduction algorithms adapt a graph embedding framework in which $\mathrm{G} = \{\mathbf{X}, \mathbf{W}\}$ is the undirected weighted graph and $\mathbf{W}$ is the $n \times n$ data dependent similarity or *affinity* matrix. The algorithms utilize the notion of affinity weights $\mathbf{W}_{ij} \in [0, 1]$ to measure the "distance" between two sample observations. The affinity functions do not utilize class label information, but rather characterizes the neighborhood relationships between all pairs of points based on feature differences. A popular approach to measure the affinity between samples $\mathbf{x}_i$ and $\mathbf{x}_j$ makes use of the *heat-kernel*,

$$\mathbf{W}_{ij} = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\gamma_i \gamma_j}\right), \tag{1}$$

where $\gamma_i = \|\mathbf{x}_i - \mathbf{x}_i^{(k_{\text{nn}})}\|$ denotes the local scaling of data samples in the neighborhood of $\mathbf{x}_i$, and $\mathbf{x}_i^{(k_{\text{nn}})}$ is the k_{nn}-nearest neighbor of $\mathbf{x}_i$. Although the heat kernel has been shown to result in effective locality preserving properties, further improvements towards sparse affinity matrices can be achieved by adapting the scaling parameter γ_i to the local data statistics which often provide a stronger adaptivity to the underlying structure of the embedded image manifolds. The affinity matrix can be modified to include spatial context, which can have a significant impact for manifold learning with hyperspectral images, as discussed in a later section.

When considering multiple data sources (e.g. co-registered gridded imagery data), disparity in the resulting feature spaces can be addressed via separate affinity matrices dedicated to each source. In the realm of kernel methods, a simple approach that has been exploited for geospatial image analysis utilizes composite kernels (e.g. a weighted linear mixture of kernels, each dedicated to a data source: $\mathbf{W}_{ij} = \sum_{k=1}^K \alpha_k \mathbf{W}_k(\mathbf{x}_i^k, \mathbf{x}_j^k)$, s.t $\alpha_k \geq 0$ and $\sum_{k=1}^K \alpha_k = 1$) to create a unified Gram matrix that characterizes relations across different input sources [23]. In the context of manifold learning algorithms, such an approach is particularly relevant for algorithms that operate directly on the affinity matrix, $\mathbf{W}$. Various complex functional forms for $\mathbf{W}_k$ can be adapted, although the heat-kernel defined in equation (1) remains a popular choice.

Dimensionality Reduction via Graph Laplacian of Spectral Features

Nonlinear manifold learning methods are broadly characterized as global or locally based approaches, and often represented using a graph embedding framework [13]. Global manifold methods retain the fidelity of the overall topology of the data set, but have greater computational overhead for large data sets, while local methods preserve local geometry and are computationally efficient because they only require sparse matrix computations. Although global manifolds seek to preserve geometry across all scales of the data and have less tendency to overfit, which is beneficial for generalization in classification, local methods may yield good results for data sets which have significantly different sub-manifolds.

Many popular existing approaches involve models that compute embeddings to preserve pairwise distances, seeking the global structure of data based on local linear fits. Manifold learning algorithms such as isometric feature mapping (ISOMAP) [2], kernel principal component analysis (KPCA) [3], and locally linear embedding (LLE) [4], for example, have received much attention because of their firm theoretical foundation associated with the kernel and eigenspectrum framework.

In general, given a data matrix $\mathbf{X}$, the dimensionality reduction problem[1] seeks to find a set of manifold coordinates $\mathbf{Y} = \{\mathbf{y}_i\}_{i=1}^n, \mathbf{y}_i \in \mathbb{R}^p$, where typically, $m \ll p$, through a feature mapping $\Phi : \mathbf{x} \rightarrow \mathbf{y}$, which may be analytical (explicit) or data driven (implicit), and linear or nonlinear. Spectral based dimensionality reduction algorithms adapt a graph embedding platform, *i.e.* with $\mathrm{G} = \{\mathbf{X}, \mathbf{W}\}$, to compute the affinity matrix from which the graph Laplacian $\mathbf{L}$ is derived to play an important role in the framework. Here, $\mathbf{L} = \mathbf{D} - \mathbf{W}$ with a diagonal degree matrix defined by $\mathbf{D}_{ii} = \sum_j \mathbf{W}_{ij}, \forall i$.

In the one-dimensional case, where the resultant manifold coordinate for *n* samples is a vector $\mathbf{y} = [y_1, y_2, \ldots, y_n]$, the dimensionality reduction criterion for eigenspectrum based methods can be represented as

$$\mathbf{y}^* = \underset{\mathbf{y}\mathbf{B}\mathbf{y}^{\mathrm{T}}=1}{\operatorname{argmin}} \sum \|y_i - y_j\|^2 \mathbf{W}_{ij} \tag{2}$$

$$= \underset{\mathbf{y}\mathbf{B}\mathbf{y}^{\mathrm{T}}=1}{\operatorname{argmin}} \mathbf{y}\mathbf{L}\mathbf{y}^{\mathrm{T}} \tag{3}$$

where $\mathbf{B}$ is a constraint matrix that depends on the formulation of the dimensionality reduction method. In many algorithms the constraint removes any arbitrary scaling factors in the embedding space. For example, setting $\mathbf{B}$ to a diagonal matrix often yields the required scale normalization. Table I summarizes various constraints that are encountered with traditional and modern graph embedding algorithms. The underlying goal is for sample pairs of larger weight to have manifold coordinates that are closer to each other, under a unique data geometry characterized by the graph Laplacian $\mathbf{L}$. The solution of the optimization problem can be obtained by solving the eigen-decomposition problem $\mathbf{L}\mathbf{y} = \lambda\mathbf{B}\mathbf{y}$, where the one-dimensional manifold coordinates $\mathbf{y}$ are given by the eigenvector with the smallest non-zero eigenvalue. This one-dimensional case can be easily generalized to the multi-dimensional case through the following expansion

[1]For the hyperspectral dataset used in this paper, the "optimal" dimensionality is found to be approximately 8 for the classical global manifold learning embeddings, 15-17 for the local embeddings, and 8-10 for the iterative embeddings.

$$\mathbf{Y}^* = \underset{\mathbf{YBY}^{\mathrm{T}}=\mathbf{I}}{\operatorname{argmin}} \operatorname{tr}(\mathbf{YLY}^{\mathrm{T}}) \quad (4)$$

where $\mathbf{I}$ is the identify matrix. Analogous to the one-dimensional case, the manifold coordinates $\mathbf{Y}$ of target dimension *p* can be obtained from the eigenvectors corresponding to the *p* smallest non-zero eigenvalues. Each of the kernel based manifold learning algorithms summarized here can be described in terms of this common framework with different Laplacian matrices and constraints. For a detailed discussion, see [13].

ISOMAP and Kernel PCA are the most widely applied global manifold learning approaches for nonlinear dimensionality reduction. The ISOMAP method assumes that the local feature space formed by the nearest neighbors is linear, and the global nonlinear transformation can be found by connecting these piecewise linear spaces [2]. Defining $\mathbf{X}_i$, the set of neighborhood nodes of node $\mathbf{x}_i$, a distance matrix $\mathbf{S}'$, is computed whereby the Euclidean distance to node $\mathbf{x}_j \in \mathbf{X}_i$ is computed, and the distance beyond $\mathbf{X}_i$ is accumulated along the shortest path to obtain a shortest path network $\mathbf{S}_{\mathrm{stp}}$. Dimensionality reduction is then accomplished through multidimensional scaling (MDS). The computational burden of computing the geodesic distance matrix scales as $O(n^2 \log n)$, motivating development of approximation methods such as Landmark ISOMAP (L-ISOMAP). These methods avoid the computation for the kernel matrix by selecting a subset of the original points, referred to as "landmark samples", for which the geodesic distance computation is performed and the remainder of the points are inserted into the "backbone", thereby reducing the computational cost of the method to $O(\ell n \log n)$, where ℓ is the number of landmark samples [24], [25]. Kernel PCA is a nonlinear extension of linear PCA in a feature space induced by a kernel function [3].

Local kernel based manifold learning methods include locally linear embedding (LLE) [4], local tangent space alignment (LTSA) [7] and Laplacian eigenmaps (LE) [8]. All three methods are initiated by constructing a nearest neighborhood for each data point, and the local structures are then used to obtain a global manifold. According to the framework, by solving the eigenvalue problem $\mathbf{LY} = \lambda\mathbf{BY}$, the embedding $\mathbf{Y}$ is provided by the eigenvectors corresponding to the 2$\sim$(*p*+1) smallest eigenvalues (the eigenvector that corresponds to the smallest zero eigenvalue is a unit vector with equal elements and is discarded). In LLE [4], the local properties of each neighborhood are represented by the linear coefficients that best reconstruct each data point from its neighbors. In LTSA [7], the local geometry is described by the local tangent space of each data point, and the global manifold is determined by aligning the overlapping local tangent spaces. In LE [8], the weighted neighborhood graph of each data point is obtained by calculating the pairwise distances between neighbors, where the distance is normally calculated using a Gaussian kernel function with parameter σ. The embeddings are obtained by minimizing the total distance between each data point and its neighbors in the low dimensional space. Parameter settings, including the size of the neighborhood, for both global and local manifold learning methods, and intrinsic dimensionality are selected experimentally and are usually robust over a range of values.

Supervised implementations of local manifold learning have also been developed for classification. Unsupervised local manifold learning approaches search the k spectral neighbors of a given point, whereas supervised local manifold learning approaches identify only the neighbors that are of the same class as the given point, often making these methods more attractive for classification [26], [27]. Supervised local manifold learning approaches then map all the training data from the same class onto a single point in the embedded space, resulting in computational complexity of $O(mn_1n_2)$, where n_1 and n_2 represent the number of training and testing samples respectively. Assuming there are c classes, the outputs are c orthogonal vectors $\mathbf{Y}^c = \left[\mathbf{y}^1, \cdots, \mathbf{y}^c\right] \in \mathbb{R}^{p\times c}$. The kernel out-of-sample extension method is attractive for unsupervised kernel-based embedding of large data sets, but is required for testing data when training data are embedded via supervised local manifold learning methods [27].

New Directions in Manifold Learning

Manifold Learning for Multi-Temporal Image Data

Classification of remotely sensed data from multiple scenes acquired at different times or from spatially disjoint areas is an important problem where it is often desirable to exploit labeled data from one time or area to classify data from a different time or area. Although global manifolds are assumed to be similar, spectral shifts in classes over space or time typically manifest themselves as localized variations in the manifold. When the goal is to exploit limited labeled data in a transfer learning mode to classify data in other scenes, changes in the manifold between images can result in misclassification of similar classes. Recent investigations that seek to jointly exploit the global and local characteristics of images [28], [29] and manifold alignment [30], [31] provide the foundation for a correspondence based framework to classify hyperspectral data acquired in multiple time periods [32], or from spatially disjoint areas.

In [33] a joint manifold over time periods T_1 and T_2 was obtained using the distance matrix

$$\mathbf{W}_G = \begin{bmatrix} \mathbf{W}_{\mathbf{x}^{T_1},\mathbf{x}^{T_1}} & \mathbf{W}_{\mathbf{x}^{T_1},\mathbf{x}^{T_2}} \\ \mathbf{W}_{\mathbf{x}^{T_2},\mathbf{x}^{T_1}} & \mathbf{W}_{\mathbf{x}^{T_2},\mathbf{x}^{T_2}} \end{bmatrix}, \quad (5)$$

where $\mathbf{W}_{\mathbf{x}^{T_1},\mathbf{x}^{T_1}}$ and $\mathbf{W}_{\mathbf{x}^{T_2},\mathbf{x}^{T_2}}$ are geodesic distances between points within the two images (intra-image distances) which capture the global geometry of the data manifolds,

and $\mathbf{W}_{\mathbf{x}^{T_1},\mathbf{x}^{T_2}}$ and $\mathbf{W}_{\mathbf{x}^{T_2},\mathbf{x}^{T_1}}$ represent the connection between the two images (inter-image distance). The inter-image distances and the resulting alignment are based on u corresponding pairs $\left(\mathbf{x}_{c_p^i}^{T_1}, \mathbf{x}_{c_q^i}^{T_2}\right)$, $i \in [1, u]$, determined from the spatial-spectral optimization

$$\arg\min_{p\leq n_1, q\leq n_2} \left(\left\|\mathbf{x}_p^{T_1} - \mathbf{x}_q^{T_2}\right\| + a\left\|\mathbf{s}_p^{T_1} - \mathbf{s}_q^{T_2}\right\|\right), \quad (6)$$

where $\left\{\mathbf{s}_1^{T_1}, \mathbf{s}_2^{T_1}, \cdots, \mathbf{s}_{n_1}^{T_1}\right\} \in \mathbb{R}^2$, and $\left\{\mathbf{s}_1{}^{T_2}, \mathbf{s}_2^{T_2}, \cdots, \mathbf{s}_{n_2}^{T_2}\right\} \in \mathbb{R}^2$ are spatial coordinates of the pair of images with n_1 and n_2 pixels, respectively. Distances between points on the two manifolds are defined in terms of distances to corresponding pairs within the respective manifold: $\mathbf{W}_{\mathbf{x}_i^{T^1},\mathbf{x}_j^{T^2}}(i,j) = \min\left(\mathbf{W}_{\mathbf{x}_i^{T_1},\mathbf{x}_{c_p^i}^{T_1}} + \mathbf{W}_{\mathbf{x}_j^{T_2},\mathbf{x}_{c_q^i}^{T_2}}\right)$, thereby preserving local relations between arbitrary points and their nearest corresponding pair. The optimal joint manifold feature space $\mathbf{Y} = \{\mathbf{y}_1, ..., \mathbf{y}_{n_1}, \mathbf{y}_{n_1+1}, ..., \mathbf{y}_{n_1+n_2}\} \in \mathbb{R}^p$, $p << m$ is computed by minimizing the cost function $E = \|\tau(\mathbf{W}_\mathbf{X}) - \tau(\mathbf{W}_\mathbf{Y})\|$ where $\mathbf{W}_\mathbf{Y}$ is a distance matrix with elements $\mathbf{W}_\mathbf{Y}(i,j) = \sqrt{(\mathbf{y}_i - \mathbf{y}_j)}$, and the τ operator converts distance that characterizes geometry to inner products. The resulting problem is solved using classical multidimensional scaling, yielding the respective eigenvectors. In recent work, Tuia *et al*, [34] also utilized manifold alignment in conjunction with linear, invertible projections to jointly exploit and synthesize data from multiple sensors.

Locality Preserving Discriminative Dimensionality Reduction

Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), and their many variants, such as subspace LDA, stepwise LDA [35], [36], etc. are commonly used for feature extraction prior to classification of hyperspectral data. Under the assumption of homoscedastic Gaussian class-conditional distributions, LDA is optimized for classification tasks, but does not perform well when the data are heteroscedastic Gaussian, and can fail for non-Gaussian data. This makes such projections inappropriate for Bayesian classifiers relying on Gaussian Mixture Models (GMMs), or for classifiers that assume the decision surfaces to be substantially nonlinear (e.g. nonlinear Support Vector Machines in a kernel-induced space). This issue is particularly relevant for hyperspectral imagery, where several factors can lead to deviation from such assumptions, including variable illumination conditions, significant mixing between the target pixel and background.

Local Fisher's Discriminant Algorithm (LFDA) [37] was developed as an extension to LDA to accommodate class distributions that are not uni-modal homoscedastic Gaussian, combining the discriminative properties of LDA with properties of unsupervised locality-preserving projections (LPP) [38]. Unlike LDA or PCA, LPP seeks to find a linear map that preserves the local-neighborhood structure of the data in the projected subspace — i.e., neighborhood points in the original input space remain neighbors in the LPP-embedded space, and vice-versa. LFDA obtains good between-class separation in the projection while preserving the within-class local structure [37]. It can hence be expected that LFDA should be a useful feature reduction algorithm for supervised classification tasks, particularly for problems where local structures convey relevant information (e.g. when the data lie on a complex manifold in the input space) and need to be preserved. In recent work for supervised hyperspectral image analysis tasks [14], [15], [39], [40], LFDA and its variants have been found to be very effective feature extraction algorithms, particularly when paired with powerful Bayesian classifiers, such as Gaussian Mixture Models. The heat kernel's normalized version has been adapted for LFDA to compute the *local* between-class $\mathbf{W}_{ij}^{(\text{lb})}$ and within-class $\mathbf{W}_{ij}^{(\text{lw})}$ weights as defined by,

$$\mathbf{W}_{ij}^{(\text{lb})} = \begin{cases} \mathbf{W}_{ij}(1/n - 1/n_l), & \text{if } z_i = z_j = l, \\ 1/n, & \text{if } z_i \neq z_j, \end{cases} \quad (7)$$

$$\mathbf{W}_{ij}^{(\text{lw})} = \begin{cases} \mathbf{W}_{ij}/n_l, & \text{if } z_i = z_j = l, \\ 0, & \text{if } z_i \neq z_j. \end{cases} \quad (8)$$

Here n_l is the number of available training samples for the l^{th} class, $\sum_{l=1}^{c} n_l = n$ and class labels are denoted by $z_i \in \{1, 2, \ldots, c\}$, where c is the number of classes.

In LFDA, the *local* between-class $\mathcal{S}^{(\text{lb})}$ and within-class $\mathcal{S}^{(\text{lw})}$ scatter matrices are defined as

$$\mathcal{S}^{(\text{lb})} = \frac{1}{2}\sum_{i,j=1}^{n} W_{ij}^{(\text{lb})}(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^\top, \quad (9)$$

$$\mathcal{S}^{(\text{lw})} = \frac{1}{2}\sum_{i,j=1}^{n} W_{ij}^{(\text{lw})}(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^\top, \quad (10)$$

LFDA seeks to find a projection Φ_{LFDA} that maximizes the "local" Fisher's ratio as defined using the local scatter matrices defined above. The solution is obtained by solving the generalized eigenvalue problem $\mathcal{S}^{(\text{lb})}\Phi_{\text{LFDA}} = \Lambda\mathcal{S}^{(\text{lw})}\Phi_{\text{LFDA}}$, where Λ is the diagonal eigenvalue matrix.

Note that based on (9), and (10), LFDA can be thought of as a "localized variant" of LDA, since it ensures that local neighborhood structures are preserved by incorporating an appropriate scaling of the scatter matrices. Hence, when the data in the input space lie on nonlinear manifolds, or in general, possess non-Gaussian, even multi-modal class-conditional statistics, LFDA is expected to outperform traditional linear projection based dimensionality reduction approaches. Another benefit of scaling the LFDA based scatter matrices is that the between-class scatter matrix is no-longer rank-limited to $c-1$. Thus, the "optimal" dimensionality of the projected subspace is no-longer restricted to $c-1$.

Although LFDA serves as an effective feature reduction strategy for hyperspectral images, it is also prone to statistical ill-conditioning when the training sample size is small.

In recent work [41], a segmented feature reduction approach was developed wherein the high-dimensional hyperspectral space is partitioned into contiguous subspaces, followed by LFDA based feature reduction and Gaussian Mixture Model based classification. Hyperspectral imagery exploits such an approach naturally, since the correlation structure of the spectral feature space is often strongly block-diagonal (nearby bands are much more correlated than bands farther apart). The resulting approach showed substantial robustness to the small-sample-size problem. Other approaches to discriminative feature reduction inspired by manifold learning are also emerging for hyperspectral image analysis. For example, in [42], a nearest feature line embedding transformation is proposed for hyperspectral dimensionality reduction, that also seeks to preserve the local manifold structure under the embedding.

Manifold Learning for Spatial-Spectral Classification of HSI

Traditional nonlinear dimension reduction approaches treat samples as statistically independent, ignoring the local spatial relationships among pixels that occur in patches, as well as the spatially disjoint locations of many spectrally similar classes. Spatial issues have been addressed in many ways by the image processing and remote sensing communities, including Markov Random Fields, vectors with stacked spectral-spatial features, morphological profiles, and segmentation (See [43] for a comprehensive review).

Recent work related to feature extraction from hyperspectral data has also addressed local spatial relationships via composite and other combined kernels [23], [44]–[46], tensor embedding [47], and iterative methods [22], [48], [49]. Forero and Manian [50] proposed nonlinear diffusion partial differential equations (PDEs) for spatial preprocessing of hyperspectral images, and the results demonstrated a significant improvement in classification performance. Represented in the context of affinities, HSI spectral and spatial neighborhood relations $\mathbf{W}_{ij} = W(\mathbf{s}_i, \mathbf{s}_j, \mathbf{x}_i, \mathbf{x}_j)$ can be computed through a weighted kernel function

$$W(\mathbf{s}_i, \mathbf{s}_j, \mathbf{x}_i, \mathbf{x}_j) = \exp\left\{\frac{-\|\mathbf{s}_i - \mathbf{s}_j\|^2}{\sigma_s^2}\right\} \cdot \tilde{W}_p(\mathbf{x}_i, \mathbf{x}_j) \quad (11)$$

where $\mathbf{s}_i$ denotes the spatial coordinates of pixel i, $\mathbf{x}_i$ denotes the spectral m-dimensional vector. The expression $\|\mathbf{s}_i - \mathbf{s}_j\|^2$ weights image pixel values as a function of the spatial distance from the center pixel and the variance parameter σ_s and

$$\tilde{W}_p(\mathbf{x}_i, \mathbf{x}_j) = \exp\left\{-\frac{1}{2}(\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{\Sigma}^{-1}(\mathbf{x}_i - \mathbf{x}_j)\right\} \quad (12)$$

simply weights relations as a function of spectral differences between the center pixel and its neighbor pixel. With additional manipulations as shown in [22], $\tilde{W}_p(\mathbf{x}_i, \mathbf{x}_j)$ can be rewritten as

$$\tilde{W}_p(\mathbf{x}_i, \mathbf{x}_j) = \exp\left\{\frac{-n}{2} tr(\mathbf{\Sigma}^{-1}\mathbf{S})\right\} \quad (13)$$

where $\mathbf{S}$ is the sample covariance. $\mathbf{\Sigma}^{-1}$ can be obtained by seeking an orthogonal decomposition of the true covariance matrix $\mathbf{\Sigma}$. This can be achieved through an efficient, robust sparse matrix transform [48], [51] to decorrelate HSI bands. The resulting affinity function infuses local adaptivity and spatial sensitivity to the neighborhood graph, which leads to preservation of local disjoint neighborhoods that are compact and similar, benefiting hyperspectral data embedding. Fig. 2 depicts the eigen-spectra corresponding to the spatially weighted Laplacian graph for a hyperspectral image. The uniqueness and smoothness of eigenvalues demonstrate the ability of affinity functions to capture both local and global structures in the data. Smooth, rapidly decaying eigenvalues suggest a neighborhood graph with a single, very large connected component — the case for using a heat kernel. Alternatively, smooth but slowly decaying eigenvalues suggest a neighborhood graph with various disconnected components, each based on the local spatial details of the image.

Iterative Graph Embedding for Dimensionality Reduction

Nonlinear embedding formulations that ignore spatial relationships often collapse maps towards the center coordinates of the embedding space, thereby increasing the crowding or overlapping of class boundaries. Given a spatially weighted affinity functions, high quality lower dimensional HSI visualization and improved classification performance may be achieved by adapting an iterative dynamic embedding framework.

In an iterative graph embedding framework, each affinity weight $\mathbf{W}_{ij} \in \mathbf{W}$, as computed from (11), is viewed as characterizing spring force properties between a pair of vertices i and j for all $\{(\mathbf{x}_i, \mathbf{x}_j)\} \in \mathbf{G}$. The affinities can be normalized or unnormalized for each observed pixel pair in $\mathbf{X}$. The embedding of $\mathbf{G}$ can then be interpreted as an assignment of positions to vertices in a p-dimensional space $\mathbb{R}^p$. With the notation $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_n\}$ denoting the assigned embedding of $\mathbf{G}$, where $\mathbf{y}_i \in \mathbb{R}^p$ is the position of the map of the ith vertex. An optimal embedding $\mathbf{Y}$ can be obtained through an iterative optimization scheme whose goal is to establish the minimum energy configuration state of $\mathbf{G}$. The quality of the embedding representation is heavily dependent on the choice of both the objective function and the kernel function used to compute the affinity matrix $\mathbf{W}$.

Iterative embedding of data is based on an intuitive premise. Assume that $\mathbf{y} = \left\{\mathbf{y}_1^T, \mathbf{y}_2^T, \cdots, \mathbf{y}_n^T\right\}^T$ is a vector in $\mathbb{R}^{Np}$ that denotes the state of $\mathbf{G}$. The framework builds on a dynamic model formulation [52], to employ pair-wise distance dependent functions and a neighborhood characterized graph to control the grouping of similar vertex maps. The dynamics evolve in continuous time; as such, the velocity as determined by the additive group effect on each vertex i,

and at position $\mathbf{y}_i$ is described by

$$\dot{\mathbf{y}}_i = \sum_{j \neq i} (\mathbf{y}_i - \mathbf{y}_j)\{F_r^{ij}(\|\mathbf{y}_i - \mathbf{y}_j\|) - F_a^{ij}(\|\mathbf{y}_i - \mathbf{y}_j\|)\} \quad (14)$$

$F_r^{ij} : \mathbb{R}^+ \to \mathbb{R}^+$ denotes the repulsion term for dispersing all embedding pixel maps, whereas $F_a^{ij} : \mathbb{R}^+ \to \mathbb{R}^+$ represents the attraction term for similar pixels.

Functional forms are selected such that an attraction term dominates the pair-wise interaction between vertex maps at large distances, while at short distances the repulsion term dominates, and in-between there is a unique distance ϵ_{ij} at which both terms will be in equilibrium - defining a minimum energy configuration state and hence an optimal positioning of pairwise vertex maps. To generate the corresponding force field, the framework assumes the existence of pair-wise dependent functions $U_{att}^{ij} \to \mathbb{R}^+ \to \mathbb{R}^+$ and $U_{rep}^{ij} \to \mathbb{R}^+ \to \mathbb{R}^+$ such that $\nabla_{\mathbf{y}_i} \mathbf{W}_{ij} U_{att}^{ij}(\|\mathbf{y}_i - \mathbf{y}_j\|) = F_a^{ij}(\|\mathbf{y}_i - \mathbf{y}_j\|)(\mathbf{y}_i - \mathbf{y}_j)$ and $\nabla_{\mathbf{y}_i} U_{rep}^{ij}(\|\mathbf{y}_i - \mathbf{y}_j\|) = F_r^{ij}(\|\mathbf{y}_i - \mathbf{y}_j\|)(\mathbf{y}_i - \mathbf{y}_j)$, where U_{att}^{ij} and U_{rep}^{ij} are viewed as artificial attraction and repulsion potential energy functions that determine the trajectories of vertex maps. The general embedding framework based on the dynamic model has the form

$$\dot{\mathbf{y}}_i = \sum_{j \neq i} \nabla_{\mathbf{y}_i} \{U_{rep}^{ij}(\|\mathbf{y}_i - \mathbf{y}_j\|) - \mathbf{W}_{ij} U_{att}^{ij}(\|\mathbf{y}_i - \mathbf{y}_j\|)\} \quad (15)$$

Following the negative gradient, *i.e.* to achieve an equilibrium state for (15), attraction and repulsion potential functions should be chosen such that the minimum of $U_{att}^{ij}(\|\mathbf{y}_i - \mathbf{y}_j\|)$ occurs around $\|\mathbf{y}_i - \mathbf{y}_j\| = 0$, whereas the minimum of $-U_{rep}^{ij}(\|\mathbf{y}_i - \mathbf{y}_j\|)$ occurs around $\|\mathbf{y}_i - \mathbf{y}_j\| \to \infty$, and that the minimum of the interactive field $U_{att}^{ij}(\|\mathbf{y}_i - \mathbf{y}_j\|) - U_{rep}^{ij}(\|\mathbf{y}_i - \mathbf{y}_j\|)$ occurs at $\|\mathbf{y}_i - \mathbf{y}_j\| = \epsilon_{ij}$, thus defining the equilibrium state of dynamic model. This general framework exhibits strong unifying properties that are applicable for deriving novel iterative multidimensional artificial field embedding algorithms. Further illustrations in this study demonstrate its use for interpreting some of the existing nonlinear dimensionality reduction models, *e.g.* reformulation of the stochastic neighbor embedding [53].

Multidimensional Artificial Field Embedding

Following the criteria described in the previous section, an attraction term according to a quadratic form can be chosen, *i.e.* $U_{att}^{ij}(\|\mathbf{y}_i - \mathbf{y}_j\|) = \xi_a \|\mathbf{y}_i - \mathbf{y}_j\|^2$. The notion of a repulsion force can be interpreted as a barrier constraint that can be captured by an indicator function, even though its gradient is difficult to compute. There are continuous indicator function approximations that yield useful repulsion terms as summarized in Table I. An effective repulsion potential function used here has the form $U_{rep}^{ij}(\|\mathbf{y}_i - \mathbf{y}_j\|) = \frac{\xi_r}{\|\mathbf{y}_i - \mathbf{y}_j\|^2}$. The parameters ξ_a and ξ_r reflect the attraction and repulsion force magnitude. Combining the two terms yields a multidimensional artificial field embedding unbounded repulsion model (MAFE-UR) [48],

$$U(\mathbf{y}) = \sum_{i=1} \sum_{j \neq i} \left\{ \xi_a \|\mathbf{y}_i - \mathbf{y}_j\|^2 \mathbf{W}_{ij} - \frac{\xi_r}{\|\mathbf{y}_i - \mathbf{y}_j\|^2} \right\} \quad (16)$$

Obtaining the optimal embedding maps involves solving a non-convex optimization problem, $\operatorname{argmin}_{\mathbf{y} \in \mathbb{R}^{Np}} U(\mathbf{y})$, whose solution space is known to exhibit many local minima and instabilities for a standard gradient descent algorithm. A much improved stable and efficient iterative updating scheme can be devised in the form of a local adaptive stochastic descent framework,

$$\mathbf{y}^{(t+1)} = \mathbf{y}^{(t)} - \alpha^{(t)} \nabla U(\mathbf{y}^t) \quad (17)$$

to yield the optimal maps. Where $\alpha^{(t+1)} = \alpha^{(t)} + \gamma_1 \langle \nabla U(\mathbf{y}^{(t-1)}), \nabla U(\mathbf{y}^{(t)}) \rangle + \gamma_2 \langle \nabla U(\mathbf{y}^{(t-2)}), \nabla U(\mathbf{y}^{(t-1)}) \rangle$ is the common adaptive learning rate. γ_1 and γ_2 are the meta-learning rates. This adaptation scheme exploits gradient-related information from the current as well as the two previous embedding coordinates in the sequence to introduce stability. The computational burden of computing the gradient scales as $O(n^2)$, motivating the need to develop faster approximation methods or finding a closed form solution to $\operatorname{argmin}_{\mathbf{y} \in \mathbb{R}^{Np}} U(\mathbf{y})$.

Stochastic Neighbor Embedding

Hinton and Roweis [53] developed a stochastic neighbor embedding (SNE) method for preserving neighbor relations based on probabilities in the lower dimensional space. The original SNE method assumes that edge weights are anti-symmetric Gaussian probabilities $\mathbf{W}_{ij}$ (i.e. $\mathbf{W}_{ij} \neq \mathbf{W}_{ji}$) of pairs of vertices being neighbors in the higher dimensional space. Considering a symmetric version of $\mathbf{W}_{ij}$ the high dimensional probability edge weights are defined using the Gaussian functions of the form

$$\mathbf{W}_{ij} = \frac{\exp\{-\|\mathbf{x}_i - \mathbf{x}_j\|^2 / 2\sigma_i\}}{\sum_{r \neq i} \exp\{-\|\mathbf{x}_r - \mathbf{x}_i\|^2 / 2\sigma_i\}} \quad (18)$$

where σ_i is computed using a binary search method ensuring that the entropy of the distribution $\mathbf{W}_i$ is approximately $\log(k)$, where k is the effective number of neighbors. In the lower dimensional space, SNE assumes symmetric Gaussian probabilities $\tilde{\mathbf{W}}_{ij}$ between embedding coordinates, *i.e.* embedding graph weights are computed as

$$\tilde{\mathbf{W}}_{ij} = \frac{\exp\{-\|\mathbf{y}_i - \mathbf{y}_j\|^2\}}{\sum_{r \neq i} \exp\{-\|\mathbf{y}_r - \mathbf{y}_i\|^2\}} \quad (19)$$

SNE proceeds to compute for the maps by minimizing a sum of Kullback Leibler(KL) objective functions

$$\sum_i KL(\mathbf{W}_i || \tilde{\mathbf{W}}_i) = \sum_i \sum_{j \neq i} \mathbf{W}_{ij} \log(\frac{\mathbf{W}_{ij}}{\tilde{\mathbf{W}}_{ij}}) \quad (20)$$

The goal is to minimize the distortion between each of the n high dimensional neighborhood distributions $\mathbf{W}_i$'s and their

corresponding lower dimensional neighborhood distributions $\tilde{\mathbf{W}}_i$'s. The difficulty with the original formulation of SNE is encountered in the optimization algorithm, where the antisymmetric assumption poses challenges requiring many experimentally defined parameters for attaining stability. In a more recent approach, Maaten and Hinton [54] improved on SNE by prescribing a *student-t* distribution to compute the lower dimensional probabilities. The improvement led to a tSNE model that preserves meaningful structures in lower dimensional spaces. A further expansion on (20) while ignoring terms that do not depend on the unknown probabilities $\tilde{\mathbf{W}}_{ij}$, yields a functional form that makes both SNE and tSNE special cases of (15). In particular, SNE can equivalently be represented by

$$U(\mathbf{y}) = \sum_i \sum_{j \neq i} \|\mathbf{y}_i - \mathbf{y}_j\|^2 \mathbf{W}_{ij} + \log \sum_{r \neq i} \exp\{-\|\mathbf{y}_r - \mathbf{y}_i\|^2\}$$

Taking the derivative yields the gradient $\nabla U(\mathbf{y})$ that forms a dynamic equation that can be used to obtain optimal embeddings through an iterative algorithm in (17).

Spherical Manifolds and Stochastic Embedding

Other than studying manifolds on a flat surface, better visualization and increased classification performance may be achieved by seeking HSI coordinate representations on curved manifolds, which exhibit desirable properties and have been well studied in statistics [55].

To embed data onto a spherical surface one can consider a unit p-dimensional sphere to be represented as the geometric locations of all unit vectors in $\mathbb{R}^{p+1}$

$$\mathbb{S}_p = \left\{ \mathbf{y}_i \in \mathbb{R}^{p+1} : \|\mathbf{y}_i\|_2 = 1 \right\} \tag{21}$$

For every observed image pixel, the goal is to learn the optimal embedding map $\mathbf{y}_i$ and a probability distribution that preserves the neighborhood relations originating from the high dimensional space. Such a goal can be achieved by applying the SSNE framework, which when given an image $\mathbf{X}$, proceeds to compute the high dimensional symmetric probability w_{ij} that pixel i would select j as its neighbor as

$$\mathbf{W}_{ij} = \frac{W(\mathbf{s}_i, \mathbf{s}_j, \mathbf{x}_i, \mathbf{x}_j)}{\sum_{r \neq i} W(\mathbf{s}_i, \mathbf{s}_r, \mathbf{x}_i, \mathbf{x}_r)} \tag{22}$$

where W is a combined spatial-spectral kernel function. The corresponding unit spherical coordinates are obtained from using an Exit distribution [56] as a kernel density function that estimates the probability of spherical points being neighbors. The Exit distribution has the form

$$f(\mathbf{y}; \mathbf{y}_i, \rho) = \frac{1}{A_p} \frac{1 - \rho^2}{\|\mathbf{y} - \rho \mathbf{y}_i\|^p}, \quad \mathbf{y} \in \mathbb{S}_p \tag{23}$$

where A_p is the surface area of $\mathbb{S}_p$, *i.e.* $A_p = \frac{2\pi^{(p/2)}}{\Gamma(p/2)}$, $\Gamma(\cdot)$ is the Gamma function, ρ is the concentration parameter, and $\boldsymbol{y}_i$ is associated with the mean direction of the distribution. The probability $\tilde{\mathbf{W}}_{ij}$ of a spherical map i selecting map j as its neighbor is computed as

$$\tilde{\mathbf{W}}_{ij} = \frac{\|\mathbf{y}_j - \rho \mathbf{y}_i\|^{-p}}{\sum_{k \neq i} \{\|\mathbf{y}_k - \rho \mathbf{y}_i\|^{-p}\}} \tag{24}$$

An added benefit of SSNE (and other iterative embedding algorithms) is that they jointly learn the optimal low-dimensional representations and also compute probability distributions over neighborhood relations (or unnormalized relations) with the understanding that spatial proximity should play a role in establishing meaningful manifold structures. SSNE obtains an optimal embedding on a unit (hyper)sphere by iteratively solving an energy minimization problem whose cost function is defined by the sum of KL divergences between the high-dimensional distribution $\mathbf{W}_i = (\mathbf{W}_{ij})$ and the unknown spherical neighborhood distribution $\tilde{\mathbf{W}}_i = (\tilde{\mathbf{W}}_{ij})$. The optimization problem is defined as

$$\mathbf{Y}^\star = \underset{\mathbf{Y} = \left\{ \mathbf{y}_i^T \mid \mathbf{y}_i \in \mathbb{S}_p \right\}}{\operatorname{argmin}} \sum_i^n KL(\mathbf{W}_i || \tilde{\mathbf{W}}_i) \tag{25}$$

Further manipulations of (25), reveal that SSNE has a functional form of (15) applied to a constant curvature space.

Hyperspectral Image Analysis Experiments

The efficacy of manifold learning techniques for hyperspectral classification, is illustrated using the Kennedy Space Center (KSC) hyperspectral data — a standard testbed dataset, that was acquired using the NASA Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) sensor at 18-m spatial resolution. With noisy and water absorption bands removed, 176 features remain for 13 wetland and upland classes of interest. Certain KSC classes that include Cabbage Palm Hammock, and Broad Leaf/Oak Hammock upland trees; Willow Swamp, Hardwood Swamp, Graminoid Marsh, and Spartina Marsh tend to be difficult to separate in lower dimensional spaces. Their spectral signatures are mixed and often exhibit only subtle differences. Fig. 3 illustrates the dataset and ground reference information, including the total number of labeled points for each class..

Visualization of Graph Embedding

Figure 4 shows a 2-dimensional scatter plot after an ISOMAP projection for the KSC hyperspectral dataset, respectively. Similar spectral classes such as the lowland marsh grasses (Swamp, Hardwood Swamp, Graminoid Marsh, and Spartina Marsh) and the upland woodlands (Cabbage Palm Hammock, and Broad Leaf/Oak Hammock upland trees) are clustered on the manifold. Classes within the groups are difficult to separate because signatures are mixed and often exhibit only subtle differences.

Figures 5, 6, and 7 depict pixel coordinate representations after the iterative MAFE-UR, SNE, and SSNE 2-dimensional

projections, respectively. As illustrated from the embedding visualizations, both SSNE and MAFE map similar pixels onto coordinates with similar values, forming tighter disjoint clusters. The disjoint nature of embeddings is attributed to the spatial information that is captured by the dual spatial-spectral kernel function.

Classification Performance via Manifold Learning

Table II depicts the overall classification accuracy, kappa-statistic and the class-specific accuracies using unsupervised (PCA, LLE, ISOMAP) and supervised (LDA, sLLE, LFDA, SSNE) techniques, followed by a 1-NN classifier. Note that all graph based methods are sensitive to the parameter k_{nn} — the number of neighbors used when constructing the affinity matrix. However, depending upon the data (particularly it's local structure) and the embedding algorithm, the classification performance of each algorithm achieves its maximum over a range of k_{nn} values. All the labeled samples (See Fig 3) were used to develop the manifolds via unsupervised methods, and 50% of the labeled samples were used for training and 50% for testing the classifier. Random sampling was repeated 20 times, and the results represent an average accuracy over 20 trials. Manifold learning techniques outperformed PCA, provided a robust classification performance, and were particularly successful at classifying "hard" classes, such as upland vegetation classes 4, 5, and 6. The intrinsic dimensionality indicated by ISOMAP and the iterative methods was somewhat higher than for PCA, and was significantly higher for LFDA, LLE, and SLLE than for PCA. For these data and the 1-NN classifier, higher accuracies were achieved via local methods than global methods, and the value of exploiting correlation structure in the spectral data was demonstrated. The result is consistent with the work of Ma *et al* [13], where a more detailed sensitivity analysis was performed on the parameters for several global and local nonlinear manifold learning methods. Both the spectral embedding provided by iterative methods and the contribution of localized spatial information were demonstrated by the significantly higher accuracies and high quality visualizations that were achieved, although the computational overhead of such methods would need to be considered for large remotely sensed data sets.

CONCLUSIONS

Advances in hyperspectral sensing provide new capability for characterizing spectral signatures in a wide range of physical and biological systems, while inspiring new methods for extracting information from these data. Hyperspectral image data often lie on sparse, nonlinear manifolds whose geometric and topological structures can be exploited via manifold learning techniques. In this article, we focused on demonstrating the opportunities provided by manifold learning for classification of remotely sensed data. However, limitations and opportunities remain both for research and applications. Although these methods have been demonstrated to mitigate the impact of physical effects that affect electromagnetic energy traversing the atmosphere and reflecting from a target, nonlinearities are not always exhibited in the data, particularly at lower spatial resolutions, so users should always evaluate the inherent nonlinearity in the data. Manifold learning is data driven, and as such, results are strongly dependent on the characteristics of the data, and one method will not consistently provide the best results. Nonlinear manifold learning methods require parameter tuning, although experimental results are typically stable over a range of values, and have higher computational overhead than linear methods, which is particularly relevant for large scale remote sensing data sets.

Opportunities for advancing manifold learning also exist for analysis of hyperspectral and multi-source remotely sensed data. Manifolds are assumed to be inherently smooth, an assumption that some data sets may violate, and data often contain classes whose spectra are distinctly different, resulting in multiple manifolds or sub-manifolds which cannot be readily integrated with a single manifold representation. Developing appropriate characterizations that exploit the unique characteristics of these sub-manifolds for a particular data set is an open research problem, for which hierarchical manifold structures appear to have merit. To date, most work in manifold learning has focused on feature extraction from single images, assuming stationarity across the scene. Research is also needed in joint exploitation of global and local embedding methods in dynamic, multi-temporal environments and integration with semi-supervised and active learning.

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] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[3] B. Schlkopf, A. J. Smola, and K. R. Muller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Computation*, vol. 10, no. 5, pp. 583–588, 1998.

[4] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by local linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[5] V. de Silva and J. B. Tenenbaum, "Global versus local methods in nonlinear dimensionality reduction," in *Proceedings of Advanced Neural Information Processing Systems*, vol. 15, Hyatt Regency, Vancouver, B.C., Canada, 2002, pp. 713–720.

[6] L. K. Saul and S. T. Roweis, "Think globally, fit locally: unsupervised learning of low dimensional manifolds," *Journal of Machine Learning Research*, vol. 4, pp. 119–155, 2003.

[7] Z. Zhang and H. Zha, "Principal manifolds and nonlinear dimensionality reduction via local tangent space alignment," *SIAM Journal on Scientific Computing*, vol. 26, no. 1, pp. 313–338, 2004.

[8] Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Computation*, vol. 15, no. 6, pp. 1373–1396, June 2003.

[9] D. K. Agrafiotis, "Stochastic proximity embedding," *Journal of Computational Chemistry*, vol. 24, no. 10, pp. 1215–1221, 2003.

[10] S. Yan, D. Xu, B. Zhang, H. Zhang, Q. Yang, and S. Lin, "Graph embedding and extensions: A general framework for dimensionality

reduction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 1, pp. 40–51, 2007.

[11] J. He, L. Zhang, Q. Wang, and Z. Li, "Using diffusion geometric coordinates for hyperspectral imagery representation," *IEEE Geoscience and Remote Sensing Letters*, vol. 6, no. 4, pp. 767–771, 2009.

[12] A. Mohan, G. Sapiro, and E. Bosch, "Spatially coherent nonlinear dimensionality reduction and segmentation of hyperspectral images," *IEEE Geoscience and Remote Sensing Letters*, vol. 4, no. 2, pp. 206–210, 2007.

[13] M. M. Crawford, L. Ma, and W. Kim, "Exploring nonlinear manifold learning for classification of hyperspectral data," in *Optical Remote Sensing: Advances in Signal Processing and Exploitation Techniques*, S. Prasad, J. Chanussot, and L. B. (Eds), Eds. London: Springer Verlag, 2011.

[14] W. Li, S. Prasad, J. E. Fowler, and L. M. Bruce, "Locality preserving dimensionality reduction and classification for hyperspectral image analysis," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 4, pp. 1185–1198, 2012.

[15] ——, "Locality-preserving discriminant analysis in kernel-induced feature spaces for hyperspectral image classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 8, no. 5, pp. 895–898, 2011.

[16] L. Ma, M. M. Crawford, and J. W. Tian, "Anomaly detection for hyperspectral images based on robust locally linear embedding," *Journal of Infrared Millimeter and Terahertz Waves*, vol. 31, no. 6, pp. 753–762, 2010.

[17] L. Zhang, D. Tao, and X. Huang, "Sparse transfer manifold embedding for hyperspectral target detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. PP, 2013.

[18] R. Heylen, D. Burazerovic, and P. Scheunders, "Nonlinear spectral unmixing by geodesic simplex volume maximization," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 3, pp. 534–542, 2011.

[19] A. Halimi, Y. Altmann, N. Dobigeon, and J.-Y. Tourneret, "Nonlinear unmixing of hyperspectral images using a generalized bilinear model," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, no. 11, pp. 4153–4162, 2011.

[20] R. Heylen and P. Scheunders, "Calculation of geodesic distances in nonlinear mixing models: Application to the generalized bilinear model," *IEEE Geoscience and Remote Sensing Letters*, vol. 9, no. 4, pp. 644–648, 2012.

[21] J. Chi and M. M. Crawford, "Selection of landmark points on nonlinear manifolds for spectral unmixing using local homogeneity," *IEEE Geoscience and Remote Sensing Letters*, vol. 10, no. 4, pp. 711–715, 2013.

[22] D. Lunga and O. Ersoy, "Spherical stochastic neighbor embedding of hyperspectral data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, no. 2, pp. 857–871, 2013.

[23] G. Camps-Valls, L. Gomez-Chova, J. Muñoz-Marí, J. Vila-Francés, and J. Calpe-Maravilla, "Composite kernels for hyperspectral image classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 3, no. 1, pp. 93–97, 2006.

[24] Y. Chen, M. Crawford, and J. Ghosh, "Improved nonlinear manifold learning for land cover classification via intelligent landmark selection," in *IEEE International Conference on Geoscience and Remote Sensing Symposium*, 2006, pp. 545–548.

[25] V. Silva and J. Tenenbaum, *Sparse multidimensional scaling using landmark points*, June 2004.

[26] X. Geng, D. Zhan, and Z. Xhou, "Supervised nonlinear dimensionality reduction for visualization and classification," *IEEE Transactions on Systems, Man, and Cybernetics, Part B, Cybern.*, vol. 35, no. 6, pp. 1098–1107, 2005.

[27] M. Li, 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.

[28] J. Verbeek, "nonlinear image manifolds by global alignment of local linear models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 8, pp. 1236–1250, 2006.

[29] R. Wang, S. Shan, X. Chen, J. Chen, and W. Gao, "Maximal linear embedding for dimensionality reduction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 9, pp. 1776–1792, 2011.

[30] S. Lafon, Y. Keller, and R. R. Coifman, "Data fusion and multicue data matching by diffusion maps," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 11, pp. 1784–1797, 2006.

[31] C. Wang and S. Mahadevan, "Heterogeneous domain adaptation using manifold alignment," in *International Joint Conference on Artifical Intelligence (IJCAI)*, Barcelona, Spain, July 16-20 2011, pp. 1541–1546.

[32] H. L. Yang and M. M. Crawford, "Manifold alignment for multitemporal hyperspectral image classification," in *2011 IEEE Geoscience and Remote Sensing Symposium*, Vancouver, BC, July 25-29 2011, pp. 4332–4335.

[33] ——, "Learning a joint manifold with global-local preservation for multitemporal hyperspectral image classification," in *2013 IEEE Geoscience and Remote Sensing Symposium*, Melbourne, Australia, July 21-26 2013, pp. 1047–1050.

[34] D. Tuia, M. Trolliet, and M. Volpi, "Multisensor alignment of image manifolds," in *2013 IEEE Geoscience and Remote Sensing Symposium*, Melbourne, Australia, July 21-26 2013, pp. 1246–1249.

[35] S. Prasad and L. M. Bruce, "Limitations of principal component analysis for hyperspectral target recognition," *IEEE Geoscience and Remote Sensing Letters*, vol. 5, no. 4, pp. 625–629, October 2008.

[36] T. V. Bandos, L. Bruzzone, and G. Camps-Valls, "Classification of hyperspectral images with regularized linear discriminant analysis," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 47, no. 3, pp. 862–873, 2009.

[37] M. Sugiyama, "Dimensionaltity reduction of multimodal labeled data by local Fisher discriminant analysis," *Journal of Machine Learning Research*, vol. 8, no. 5, pp. 1027–1061, May 2007.

[38] X. He and P. Niyogi, "Locality preserving projections," in *Advances in Neural Information Processing System*, S. Thrun, L. Saul, and B. Schölkopf, Eds. Cambridge, MA: MIT Press, 2004.

[39] M. Cui, S. Prasad, W. Li, and L. Bruce, "Locality preserving genetic algorithms for spatial-spectral hyperspectral image classification," *IEEE Journal of Selected Topics in Applied Earth Observation and Remote Sensing*, 2013, under review.

[40] G. Zhang and X. Jia, "Feature selection using kernel based local fisher discriminant analysis for hyperspectral image classification," in *IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2011, pp. 1728–1731.

[41] S. Prasad, M. Cui, W. Li, and J. Fowler, "Segmented mixture of gaussian classification for hyperspectral image analysis," *IEEE Geoscience and Remote Sensing Letters*, 2013, to appear.

[42] Y.-L. Chang, J.-N. Liu, C.-C. Han, and Y.-N. Chen, "Hyperspectral image classification using nearest feature line embedding approach," *IEEE Transactions on Geoscience and Remote Sensing*, vol. PP, no. 99, pp. 1–10, 2013.

[43] M. Fauvel, Y. Tarabalka, J. A. Benediktsson, J. Chanussot, and J. C. Tilton, "Advances in spectral–spatial classification of hyperspectral images," *Proceedings of the IEEE*, vol. 101, no. 3, March 2013.

[44] W. Kim, M. M. Crawford, and J. Ghosh, "Spatially adapted manifold learning for classification of hyperspectral imagery with insufficient labeled data," in *IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2008, pp. 213—217.

[45] M. Fauvel, J. Chanussot, and J. A. Benediktsson, "A spatial–spectral kernel-based approach for the classification of remote-sensing images," *Pattern Recognition*, vol. 45, no. 1, pp. 381–392, 2012.

[46] H. H. Yang and M. M. Crawford, "Exploiting spectral-spatial proximity for classification of hyperspectral data on manifolds," in *Proc. IEEE Geosci. Remote Sens. Symp.*, Munich, Germany, 2012, pp. 4174–4177.

[47] L. Zhang, D. Tao, and X. Huang, "Tensor discriminative locality alignment for hyperspectral image spectral–spatial feature extraction," *IEEE Transactions on Geosciences and Remote Sensing*, vol. 51, no. 1, pp. 242–254, January 2013.

[48] D. Lunga and O. Ersoy, "Dynamic hyperspectral embedding with a spatial sensitive graph," in *Proc. IEEE Geosci. Remote Sens. Symp.*, Melbourne, AUS, 2013.

[49] ——, "Multidimensional artificial field embedding with spatial sensitivity," *IEEE Transactions on Geoscience and Remote Sensing*, vol. PP, no. 99, pp. 1–1, 2013.

[50] S. Velasco-Forero and V. Manian, "Improving hyperspectral image classification using spatial preprocessing," *IEEE TGRS*, vol. 6, no. 2,

pp. 297–301, 2009.
[51] J. Theiler, G. Cao, L. R. Bachega, and C. A. Bouman, "Sparse matrix transform for hyperspectral image processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 3, pp. 424–437, 2011.
[52] V. Gazi and K. M. Passino, "Stability analysis of swarms," *Proc. American Control Conference*, 2002.
[53] G. Hinton and S. Roweis, "Stochastic neighbor embedding," in *Proc. of ICML*, vol. 15, 2002, pp. 833–840.
[54] L. van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.
[55] K. V. Mardia and P. E. Jupp, *Directional Statistics*. New York: Wiley, 2000.
[56] S. Kato, "A distribution for a pair of unit vectors generated by brownian motion," *Bernoulli*, vol. 15, no. 3, pp. 898–921, 2009.

Author Biographies

Dalton Lunga (*dlunga@csir.co.za*) is a senior research scientist at Meraka Institute, Council for Scientific and Industrial Research, South Africa. He received a BIng degree in Electrical and Electronics Engineering from the University of Johannesburg in 2004, a Masters in Engineering from Witwatersrand University 2006, a second Masters in Electrical and Computer Engineering and a Ph.D. degree in Electrical and Computer Engineering, Purdue University in 2010 and 2012 respectively. He was a Fulbright Scholar at Purdue University, West Lafayette, IN. His research interests include statistical machine learning, signal and image processing, optimization, manifold learning, remote sensing, image reconstruction and segmentation, and data fusion.

Saurabh Prasad (*saurabh.prasad@ieee.org*) is an assistant professor of Electrical and Computer Engineering at the University of Houston. He received a B.Tech degree from Jamia Millia Islamia, New Delhi, India, in 2003, a M.S. degree from Old Dominion University, Norfolk, VA, in 2005, and a PhD degree from Mississippi State University, MS, in 2008, all in Electrical Engineering. He leads a group on geospatial image analysis at the University of Houston, and his research interests include statistical signal processing, machine learning, image processing, and data fusion.

Melba M. Crawford (*mcrawford@purdue.edu*) is the Professor of Excellence in Earth Observation at Purdue University. She received the B.S. and M.S. degrees in civil engineering from the University of Illinois at Urbana-Champaign and the Ph.D. degree in systems engineering from Ohio State University. Her research interests focus on development of advanced methods for image analysis, including: manifold learning, active learning, classification and unmixing, and application of these methods to remotely sensed data. Dr. Crawford is a Fellow of the IEEE and the current president of the IEEE Geoscience and Remote Sensing Society.

Okan Ersoy (*ersoy@purdue.edu*) is currently a Professor in the School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN. He 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. 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. Dr. Ersoy is a Fellow of the Optical Society of America, and a fellow of IEEE.

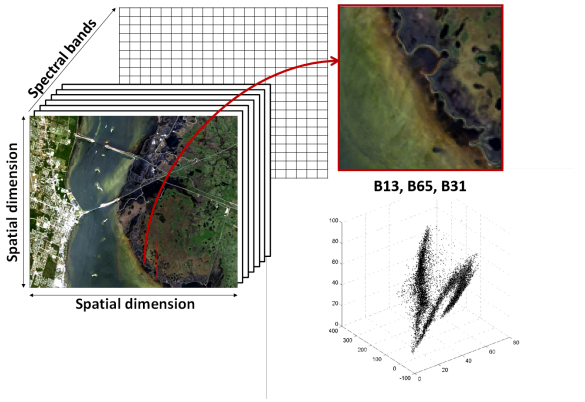

Fig. 1. True color AVIRIS hyperspectral image over Kennedy Space Center (KSC), FL. Nonlinearity in the spectral data is exhibited in a plot of bands Bands 13, 65, and 31.

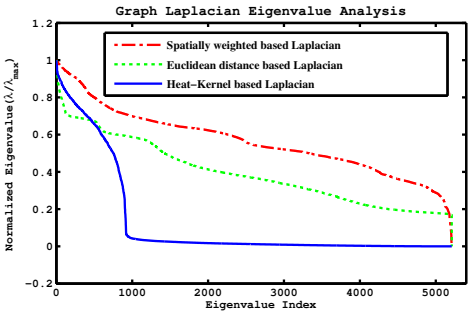

Fig. 2. Plots of normalized eigenvectors for different graph neighborhoods computed from Euclidean distance, spatially weighted, and heat kernel based Laplacian affinity functions for KSC data.

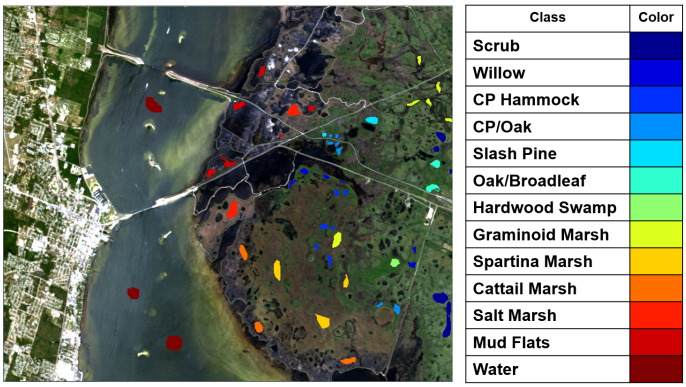

Fig. 3. Ground reference information for KSC hyperspectral data set.

Table I. Affinities And Constraints For Various Graph Embedding Algorithms

Algorithm	Affinity	Constraint	Approximation
LFDA [15]	$\mathbf{W}_{ij}^{(\mathrm{lb})}, \mathbf{W}_{ij}^{(\mathrm{lw})}$	—	none
ISOMAP [5]	$\mathbf{W}_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|^2$	$\mathbf{B} = \mathbf{I}$	none
PCA/KPCA [3]	$\mathbf{W}_{ij} = 1/n, i \neq j$	$\mathbf{B} = \mathbf{I}$	none
LLE [4]	$\mathbf{W}_{ij} = (\mathbf{M} + \mathbf{M}^T - \mathbf{M}^T\mathbf{M})_{ij}$	$\mathbf{B} = \mathbf{I}$	none
LE [8]	$\mathbf{W}_{ij} = \exp\left(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/t\right)$	$\mathbf{B} = \mathbf{D}$	none
SNE [53]	$\mathbf{W}_{ij} = \frac{\exp\{-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2\sigma_i\}}{\sum_{r\neq i}\exp\{-\|\mathbf{x}_r - \mathbf{x}_i\|^2/2\sigma_i\}}$	$I_+(f(u))$	$f(u) = \log\sum_{r\neq i} e^{-u^2}$
SSNE [22]	$\mathbf{W}_{ij} = \frac{W(\mathbf{s}_i,\mathbf{s}_j,\mathbf{x}_i,\mathbf{x}_j)}{\sum_{r\neq i} W(\mathbf{s}_i,\mathbf{s}_r,\mathbf{x}_i,\mathbf{x}_r)}$	$I_+(f(\nu))$	$f(\nu) = \log(\sum_{j=1}^{n}\nu^{-p})$
MAFE-UR [49]	$\mathbf{W}_{ij} = W(\mathbf{s}_i,\mathbf{s}_j,\mathbf{x}_i,\mathbf{x}_j)$	$I_+(f(u))$	$f(u) = \xi_r u^{-2}$

where $u = \|\mathbf{y}_i - \mathbf{y}_j\|, \nu = \|\mathbf{y}_i - \rho\mathbf{y}_j\|, f(\cdot)$ approximates the indicator constraint, $\mathbf{H} = \mathbf{I} - \mathbf{e}\mathbf{e}^T/n, \mathbf{e}$ is a n dimensional vector with $\mathbf{e} = [1, 1, \cdots, 1]^T$. $I_+(f(u)) = \begin{cases} 0, & u > \epsilon_{ij} \\ \infty, & u \leq \epsilon_{ij} \end{cases}$, ϵ_{ij} is the equilibrium point where the attraction and repulsion forces balances out. $\mathbf{T}_{ij} = [\mathbf{S}_{\mathrm{stp}}]^2, \mathbf{M} = -\mathbf{HTH}$.

Table II. The overall-accuracy (OA), Kappa-statistic, and class-specific accuracies for the 13 classes in the KSC hyperspectral dataset.

	OA	Kappa	1	2	3	4	5	6	7	8	9	10	11	12	13
PCA	88	86.7	93.8	85.8	89	62.2	50.5	44	82	86.5	95.5	92.2	95.7	87.6	99.9
ISOMAP	88.3	86.9	91.6	89.2	84.5	57.3	56.8	42.5	85.2	87.4	95.5	98.4	94.5	89.4	100
LLE	89.5	88.3	92.6	89	84	60.8	54.4	49	81.5	87.5	94.9	98.2	98.6	94.9	100
LDA	94	93.4	95.4	94	84.8	75.4	79.2	78.3	82.8	91.4	97.2	100	98.8	99.3	100
sLLE	93.2	92.4	96.4	94	93	73.1	65.5	62.3	91.7	91.3	98.9	98.6	98.5	95.4	100
LFDA	94.9	93.3	94.7	92.3	89.7	76.9	82.8	82.2	91.8	93.8	98.1	99.8	98.7	99.2	100
SNE	83.5	81.9	91.2	85.34	80.4	51.9	41.2	39.3	82.77	63.62	93.56	93.68	93.81	81.91	100
SSNE	99.42	99	98	100	100	100	95.45	100	100	100	100	100	100	98.53	100
MAFE-UR	99.6	99.72	98.3	100	100	91.21	100	86.2	100	100	100	100	98.97	100	100

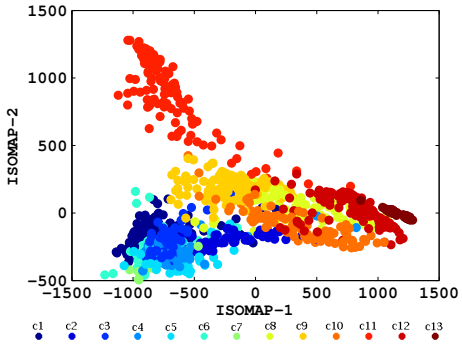

Fig. 4. 2D scatter plot of the first two dimensions of the ISOMAP embedding of KSC data.

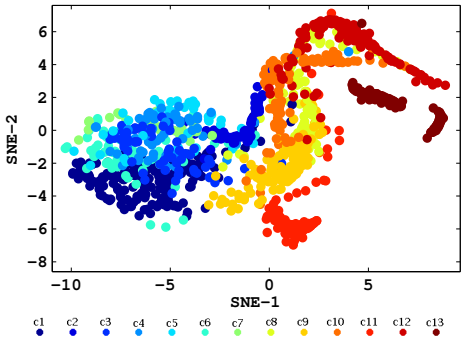

Fig. 6. 2D scatter plot of the first two dimensions of the SNE embedding of KSC data.

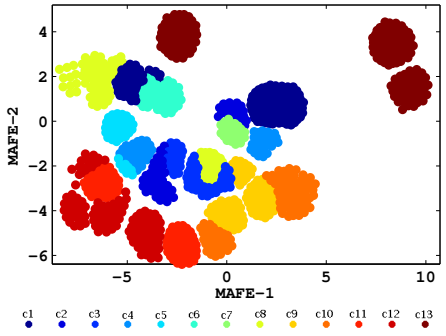

Fig. 5. 2D scatter plot of the first two dimensions of the MAFE-UR embedding of KSC data.

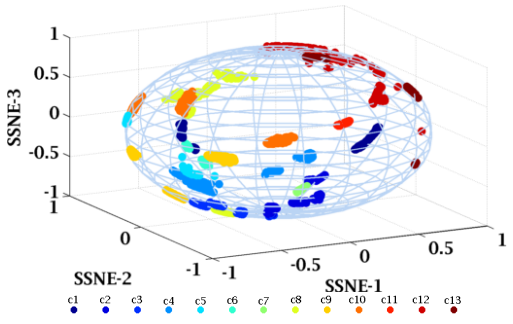

Fig. 7. 2D scatter plot of the first two dimensions of the SSNE embedding of KSC data.