Automated Estimation of Spectral Neighborhood Size in Manifold Coordinate Representations of Hyperspectral Imagery: Implications for Anomaly Finding, Bathymetry Retrieval, and Land Applications

†Charles M. Bachmann, ‡Thomas L. Ainsworth, †Robert A. Fusina

Naval Research Laboratory
Remote Sensing Division
Coastal & Ocean Remote Sensing Branch
‡Coastal Science & Interpretation Section, Code 7232
†Image Sciences Section, Code 7263
4555 Overlook Ave., SW
Washington, D.C. 20375

1. INTRODUCTION

In the past we have presented a framework for deriving a set of intrinsic manifold coordinates that directly parameterize high-dimensional data, such as that found in hyperspectral imagery [1-7]. In these previous works, we have described the potential utility of these representations for such diverse problems as land-cover mapping and in-water retrievals such as bathymetry. Because the manifold coordinates are intrinsic, they offer the potential for significant compression of the data, and are furthermore very useful for displaying data structure that can not be seen by linear image processing representations when the data is inherently nonlinear. This is especially true, for example, when the data are known to contain strong nonlinearities, such as in the reflectance data obtained from hyperspectral imaging sensors over the water, where the medium itself is attenuating [2,3,5,7]. These representations are also potentially useful in such applications as anomaly finding [2,3]. A number of other researchers have looked at different aspects of the manifold coordinate representations such as the best way to exploit these representations through the backend classifier [11], while others have examined alternative manifold coordinate models [10].

2. LOCAL DIMENSIONALITY AND NEIGHBORHOOD DEFINITION

One of the potential limits to the fidelity of these models is the issue of defining neighborhood sizes. Neighborhoods in manifold coordinate representations are the region around a sample vector point wherein the data is assumed to vary linearly. Neighborhoods must be defined in order to determine the nonlinear (geodesic) distances between spectral samples over the hypersurface of the high-dimensional data via a shortest path algorithm. These nonlinear distances become the inputs to a geodesic distance covariance matrix which is used to derive the coordinate axes of the manifold coordinate representation. Neighborhoods definitions are also crucial because they determine the extent to which the curvature of the data is modeled accurately, and most of the original approaches to choosing neighborhood sizes [12] were ad hoc (K nearest neighbors or specific radii) and, therefore, essentially user-defined by trial and error. In [3], we defined a scalable approach to deriving manifold coordinates and defined within the algorithm flow a hybrid approach to neighborhood definitions which provided a nominal search radius, and several successive search radii, while maintaining the neighborhood size between some \(K_{\text{min}}\) and \(K_{\text{max}}\). The principle reason for this approach was to accommodate the varying spectral density often in found in coastal scenes where densities can vary by as much as an order of magnitude between water and land portions of the same hyperspectral scene [3]. In [1], we proposed the idea of estimating a best choice of the neighborhood size (number of K neighbors) by looking at spectrally similar regions and demonstrated that it was possible to estimate a local intrinsic dimensionality from a log plot of the scaling relationship between the sum of distances over the k-neighbor graph as a function of sample size \(N\). The approach is based on a concept introduced by Costa and Hero in [8,9] which shows that the expected length of the distances of the K nearest neighbor graph should scale as \(N^{d-1/d}\). Log-log fits to these data for different values of \(N\) and \(K\) can then be used to identify the local intrinsic dimensionality \(d\) from the slope, and an optimal \(K\) can be selected from a graph of the estimated local dimensionality as a function of \(K\).
3. IMPLEMENTATION AND RESULTS

In this paper, we ensure that the regions where these local dimensionality estimates are made are confined to regions of a high degree of spectral similarity using the vantage point forest pre-quantization of the data that we used in [3] to identify the nearest neighbors in near linear time. The top level trees of the forest provide a natural quantization of the scene where these local dimensionality estimates can be performed. Data sets from airborne hyperspectral flights over the Virginia Coast Reserve barrier islands are used in this study to demonstrate the implications for land-cover and anomaly finding; while bathymetry retrieval from PHILLS data of the Indian River Lagoon is used to study implications of the new neighborhood for accuracy of bathymetric retrieval from hyperspectral imagery.

4. REFERENCES