Fiber Tract Clustering on Manifolds With Dual Rooted-Graphs

Andy Tsai[1,2] Carl-Fredrik Westin[1] Alfred O. Hero III[3] Alan S. Willsky[2]

[1] Dept. of Radiology
Brigham & Women's Hospital
Harvard Medical School
Boston, MA 02115

[2] Dept. of Electrical Eng.
& Computer Science
Mass. Institute of Technology
Cambridge, MA 02139

[3] Dept. of Electrical Eng.
& Computer Science
University of Michigan
Ann Arbor, MI 48109

Abstract

We propose a manifold learning approach to fiber tract clustering using a novel similarity measure between fiber tracts constructed from dual-rooted graphs. In particular, to generate this similarity measure, the chamfer or Hausdorff distance is initially employed as a local distance metric to construct minimum spanning trees between pairwise fiber tracts. These minimum spanning trees are effective in capturing the intrinsic geometry of the fiber tracts. Hence, they are used to capture the neighborhood structures of the fiber tract data set. We next assume the high-dimensional input fiber tracts to lie on low-dimensional non-linear manifolds. We apply Locally Linear Embedding, a popular manifold learning technique, to define a low-dimensional embedding of the fiber tracts that preserves the neighborhood structures of the high-dimensional data structure as captured by the method of dual-rooted graphs. Clustering is then performed on this low-dimensional data structure using the k-means algorithm. We illustrate our resulting clustering technique on both synthetic data and on real fiber tract data obtained from diffusion tensor imaging.

1. Introduction

Diffusion tensor imaging is an emerging MRI-based technology designed to measure the diffusivity of the water molecules in local tissue beds. Of particular interest is the application of this technique to the brain parenchyma. Specifically, by taking advantage of the property that water molecules diffuse preferentially along the length of the axonal tracts and less so in the direction perpendicular to the tracts, the white matter fiber structures can be characterized and their connectivity mapped. This information can then be organized for use in surgical planning and in studying a variety of disorders including neurodegenerative diseases, addiction, epilepsy, and mental disorders. A good review of diffusion magnetic resonance imaging can be found in [11].

Accurate and efficient visualization of this large and complicated 3D fiber tract data set to gain clinical insight is extremely difficult. Motivated by this shortcoming, this paper focuses on the problem of clustering the fiber tracts into natural anatomical bundles for ease of display and to facilitate information exchange and interpretation.

1.1. Previous Works

Over the past few years during the time when diffusion tensor imaging modality has gained immense popularity, fiber tract clustering has likewise gained significant attention with the development of various clustering techniques. In general, these algorithms all share the common theme of first defining a similarity metric between the fiber tracts, and then employing an algorithm for clustering based on the established similarity measure. For example, a clustering algorithm similar to k nearest-neighbors approach is proposed in [6] with the similarity metric between paired fiber tracts defined in terms of the length ratio and the Euclidean distance between the corresponding segments of the paired fiber tracts. A fuzzy c-means clustering algorithm is presented in [14] incorporating various distance measures between fiber tracts including the dot product of the corresponding tangents of the tracts and the average distance between points along the tracts. In [16], an agglomerative hierarchical clustering method is used in conjunction with a distance metric based on shortest distances between points on the tracts as defined in [15]. In [5], various pairwise distances between tracts (including closest point distance, symmetric chamfer distance, and symmetric Hausdorff distance) and geometric characteristics of fibers (including length, center of mass, and second order moment) are utilized for threshold-based clustering. In [10], B-spline representations of fiber tracts are used for comparison between those fiber tracts extracted from the subject to those from an atlas, and then based on the labeled atlas of the fiber tracts, the subject's fiber tracts are clustered.

Of particular interest to this paper is the work described

in [3] which is the first to utilize manifold learning as an image processing tool for visualizing fiber tracts. Inspired by this work, three additional techniques have followed suit [2, 9, 12], and they deserve special mention as they share strong ties with the algorithm proposed in this paper. These techniques, very much like ours, employ spectral methods of various flavors for clustering with each method utilizing an affinity matrix constructed from a different fiber tract similarity measure. In [2], an Euclidean feature space (composed of the means and covariances of the points building up the fiber tracts) is used as a similarity measure for pairwise fiber tracts. Radial basis functions are employed to map this feature space to weights of an undirected graph which are then partitioned into coherent sets using the normalized cut criterion for clustering. In [9], a co-occurrence matrix containing the number of times two fibers share the same voxel is used as the affinity matrix, and eigenvalue decomposition is performed on this affinity matrix to obtain a set of eigenvectors for clustering by a k-means algorithm. In [12], a k-way normalized cut procedure is proposed for clustering with an affinity matrix composed of symmetrized Hausdorff distances between pairwise tracts.

1.2. Contributions of Our Work

In the literature, there appears to be disproportionately more effort and emphasis placed on the development of a better clustering algorithm and less so on the design of a better similarity measure between the fiber tracts, even though the latter is more important in determining the success of a fiber tract clustering algorithm. The challenge in the clustering arena is to find the most appropriate distance measure that will farthest separate the fiber tracts belonging to different clusters while keeping fiber tracts of the same cluster close by. Popular distance measures including the chamfer and the Hausdorff distances have been proposed that adequately capture the local relationship of the fiber tracts but tend to lack the ability to capture the global structure of the input data set. In this paper, we propose a similarity measure based on dual rooted diffusion [7] which provides a more geometrically descriptive measure of the similarity between fiber tracts. Importantly, it captures both the local and the global intrinsic geometry of the data set in a principled and effective manner.

Similar to the approaches taken by others [2, 9, 12], our proposed distance measure is then incorporated into a manifold clustering algorithm, which in our case is Locally Linear Embedding, for data partitioning. Manifold learning approaches seek to define a low-dimensional embedding of the input data points that preserves the neighborhood structure of the high dimensional point set. We believe that this methodology will be an effective mechanism to reveal the underlying meaningful low dimensional information hidden within high dimensional observations for successful clustering of the input fiber tract data.

1.3. Paper Organization

The rest of this paper is organized as follows. In Section 2, we introduce the similarity measure between pairwise fiber tracts that we use in our technique. Section 3 describes how we utilize a variant of the Locally Linear Embedding method for manifold clustering. Section 4 presents preliminary results of our algorithm using both synthetic and actual diffusion tensor imaging data. We offer our concluding remarks and future research directions in Section 5.

2. Fiber Tract Similarity Measure

In this section, we introduce two well studied distance measures for 3D space curves, the chamfer [1] and the Hausdorff [8] distances. Both of these distance measures provide some form of local similarity measure between fiber tracts. As a more effective means of capturing the local and global relationships between the fiber tracts, we describe how either one of these distance measures can be incorporated into a framework of dual rooted graph diffusion to obtain a novel fiber tract similarity metric which is capable of capturing the intrinsic geometry of the data set.

2.1. Chamfer and Hausdorff Distances

Let $\mathbf{X}_M = \{\mathbf{x}_1, ..., \mathbf{x}_M\}$ be the set of M fiber tracts in a high dimensional vector space $\mathbb{R}^d$. A reasonable and popular distance measure between fiber tract $\mathbf{x}_i = \{x_{i_p}\}_{p=1}^{P}$ and fiber tract $\mathbf{x}_j = \{x_{j_q}\}_{q=1}^{Q}$ is the chamfer distance which is given by the average of the distances between each point $x_{i_p} \in \mathbf{x}_i$ and its closest point in $\mathbf{x}_j$:

$$d_{chamfer}(\mathbf{x}_i, \mathbf{x}_j) = \frac{1}{n} \sum_{x_{i_p} \in \mathbf{x}_i} \min_{x_{j_q} \in \mathbf{x}_j} \|x_{i_p} - x_{j_q}\|$$

where $\|\cdot\|$ denotes the Euclidean norm. A different but equally popular distance measure between fiber tract $\mathbf{x}_i$ and fiber tract $\mathbf{x}_j$ is the Hausdorff distance [8] which is given by the *maximum*[1] of the distances between each point $x_{i_p} \in \mathbf{x}_i$ and its closest point in $\mathbf{x}_j$:

$$d_{Hausdorff}(\mathbf{x}_i, \mathbf{x}_j) = \max_{x_{i_p} \in \mathbf{x}_i} \left\{ \min_{x_{j_q} \in \mathbf{x}_j} \|x_{i_p} - x_{j_q}\| \right\}.$$

Both of these distance measures can easily be made symmetric by taking the average between $d(\mathbf{x}_i, \mathbf{x}_j)$ and $d(\mathbf{x}_j, \mathbf{x}_i)$ so as to obey the metric properties. To a certain degree, both these measures are effective in capturing the local neighborhood structures of the input fiber tract data set but inadequate in capturing distant relationships. Of note,

[1] Others have also proposed the use of the median distance instead of the average (chamfer) or the maximum (Hausdorff) distance.

an inherent problem with Hausdorff distance is that a point in $\mathbf{x}_i$ that is farthest from any point in $\mathbf{x}_j$ dominates and may inappropriately skew this distance measure.

2.2. Dual Rooted-Graphs

Motivated by the notion of a diffusion distance built upon random walks on graphs [4], a novel and more robust similarity criterion between high dimensional data points (such as fiber tracts) is introduced in [7]. This measure is suited for clustering on smooth manifolds, and is effective in capturing the intrinsic geometry of the input data.

The specifics of this algorithm based on dual rooted graphs is described here. For each fiber tract $\mathbf{x} \in \mathbf{X}_M$, recursively grow a minimum spanning tree (MST) rooted in $\mathbf{x}$ in the following manner. Start at the root node of the tree $\mathbf{x}$ at time $k = 0$ with the set $MST_0(\mathbf{x}, \mathbf{X}_M) = \mathbf{x}$. Let $MST_k(\mathbf{x}, \mathbf{X}_M)$ denote the set of fiber tracts in the tree at time k rooted at $\mathbf{x}$. With each successive discrete time point k, add a fiber tract closest to the root node $\mathbf{x}$ that is in $\mathbf{X}_M$ but not already in $MST_{(k-1)}(\mathbf{x}, \mathbf{X}_M)$. Chamfer or Hausdorff distance is used here to measure the distance between two fiber tracts in $\mathbb{R}^d$.[2] At the end of time $k = M - 1$, all the fiber tracts in $\mathbf{X}_M$ will be included in $MST_k(\mathbf{x}, \mathbf{X}_M)$ arranged in an ascending order of distance to the root node $\mathbf{x}$. Time stamps of when each fiber tract is added to the tree is also recorded. This process is repeated M times to produce M fully grown trees with each fiber tract $\mathbf{x} \in \mathbf{X}_M$ serving as a root node in this set of M MSTs.

Next, define the hitting time $\tau(\mathbf{x}_a, \mathbf{x}_b)$ between the fiber tracts $\mathbf{x}_a$ and $\mathbf{x}_b$ in $\mathbf{X}_M$ as the time k when the two MSTs rooted in $\mathbf{x}_a$ and $\mathbf{x}_b$ intersect, i.e.

$$\tau(\mathbf{x}_a,\mathbf{x}_b)=\min\{k : MST_k(\mathbf{x}_a,\mathbf{X}_M)\cap MST_k(\mathbf{x}_b,\mathbf{X}_M)\neq 0\}.$$

From an implementational stand point, the hit time $\tau(\mathbf{x}_a, \mathbf{x}_b)$ is determined by parsing $MST_k(\mathbf{x}_a, \mathbf{X}_M)$ and $MST_k(\mathbf{x}_b, \mathbf{X}_M)$ sequentially until a common point is found between the two MSTs. Once τ is found, the total path length between $\mathbf{x}_a$ and $\mathbf{x}_b$ is calculated by summing up the pairwise distances between sequential fiber tracts within each of the two MSTs up to the hit time τ. A $M \times M$ symmetric square matrix $\mathcal{T}$ containing the distances between every pairwise fiber tracts in $\mathbf{X}_m$ can be generated in this fashion. We believe that this proposed methodology captures the local structure of the input data via the chamfer or the Hausdorf distance while the global structure is captured through the complexity of the paths taken between any pair of fiber tracts in the data set via the MSTs.

3. Manifold Clustering With LLE

Various methods can be employed to analyze matrices of pairwise distances for spectral clustering, and we chose a variant of the Locally Linear Embedding (LLE) as described in [13]. The goal of LLE is to map high dimensional inputs $\mathbf{X}_M$ to low dimensional outputs $\mathbf{Y}_M$ using local linear reconstruction weights $\mathbf{W}$. To accomplish that, LLE first attempts to represent the input data manifold locally by reconstructing each data point $\mathbf{x}_i$ as weighted combination of its neighbors through the weights $\mathbf{W}$. Specifically, we seek $\hat{\mathbf{W}}$ as below:

$$\hat{W} = \arg\min_W \sum_{i=1}^{M} \|\mathbf{x}_i - \sum_{j\in\mathcal{N}(i)} W_{ij}\mathbf{x}_j\|^2 \;\text{ s.t. }\; \forall i \;\sum_j W_{ij}=1 \;.$$

However, instead of calculating the reconstruction weights based on the above equation using fiber tract data $\mathbf{x}_i$, we opt to use the method described in [13] and calculate them based on the *pairwise distances* of the fiber tracts. In particular, given the distance matrix $\mathcal{T}$ computed as described in Section 2, the nearest K neighbors $\mathcal{N}_K(i)$ of fiber tract $\mathbf{x}_i$ is identified by parsing $\mathcal{T}$ to find the K smallest non-zero elements corresponding to each fiber tract. Knowing the neighborhood structure of each fiber tract, the local covariance matrix C_{ij} of fiber tract $\mathbf{x}_i$ and its K neighboring fiber tracts $\mathbf{x}_j$ with $j \in \mathcal{N}_K(i)$ can be derived by computing the following:

$$C_{ij} = \frac{1}{2}(D_i + D_j - D_{ij} - D_0),$$

where D_{ij} is the square of the distance between the ith and the jth neighbors as provided by $\mathcal{T}$, $D_l = \sum_z D_{lz}$, and $D_0 = \sum_{ij} D_{ij}$ [13]. In terms of C_{ij}, the optimal reconstruction weights $\hat{\mathbf{W}}$ to best reconstruct each fiber tract $\mathbf{x}_i$ from its neighbors are given by:

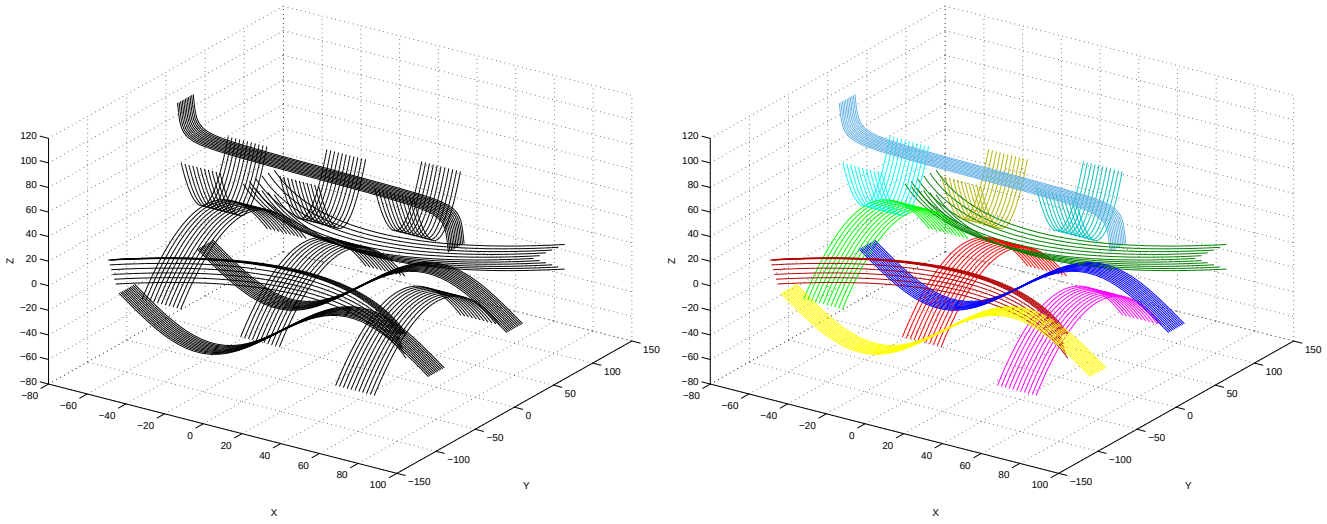
$$\hat{w}_j = \frac{\sum_k C_{jk}^{-1}}{\sum_{lm} C_{lm}^{-1}}.$$

The optimal weights $\hat{\mathbf{W}}$ is an $M \times M$ sparse matrix calculated to capture the neighborhood structure of the fiber tracts. Based on these weights, the next step is to map the high dimensional observation data $\mathbf{X}_M$ to a low dimensional vectors $\mathbf{Y}_M$ by minimizing an embedding quadratic cost functional:

$$\hat{\mathbf{Y}} = \arg\min_{\mathbf{Y}} \sum_{i=1}^{M} \|\mathbf{y}_i - \sum_{j\in\mathcal{N}(i)} \hat{W}_{ij}\mathbf{y}_j\|^2 \;\text{ s.t. }\; \begin{matrix}\mathbf{Y1} &= \mathbf{0}\\ \mathbf{YY}' &= \mathbf{I}_M\end{matrix} .$$

Importantly, only the geometric information encoded by the weights $\hat{\mathbf{W}}$ is used to construct the embedding and not the input data $\mathbf{X}_M$. Since this embedding cost functional is quadratic in $\mathbf{Y}$, it can be estimated by solving a sparse $M \times M$ eigenvector problem. The eigenvectors associated with the smallest d positive eigenvalues define the best d dimensional fit. Finally, as is common practice, a k-means method is applied to partition the resulting d eigenvectors for clustering.

[2] In fact, any reasonable distance metric between curves can be used.



(a) (b)

Figure 1. Trivial synthetic fiber tract clustering example. (a) Input data consisting of 121 synthetic fiber tracts each described by 50 data points. (b) Correct partitioning of fiber tracts into 11 bundles with $K = 4$ and $d = 10$.

4. Preliminary Results

Results based on both synthetic and real fiber tract data are presented in this section to illustrate the performance of our clustering algorithm. All algorithms are implemented in Matlab. In section 4.1, we show simulation results specifically designed to illustrate some of the features and capabilities of our algorithm as described earlier in the paper. Section 4.2 demonstrates the performance of our approach by applying it to the clustering of a real fiber tract data set obtained from diffusion tensor imaging.

4.1. Synthetic Dataset

The purpose of the first synthetic example as shown in Figure 1 is to demonstrate some of the basic features of our manifold clustering algorithm, specifically, its ability to cluster a 3D data set (121 fiber tracts each consisting of 50 data points) into multiple bundles (11) accurately and in an efficient manner (28 seconds on a 3.1GHz Intel Xeon processor). The second synthetic example shown in Figure 2 is intended to demonstrate the robustness of our clustering algorithm in a more hostile environment–one corrupted by additive noise, with complicated fiber structures, and having varying fiber tract lengths. The original synthetic data with complicated fiber tract structures is displayed in Figure 2(a). Segments of varying size fiber tracts are removed to arrive at the data set shown in Figure 2(b). Gaussian noise with zero mean and standard deviation of 1 is added to generate the input noisy test data set shown in Figure 2(c). Based on this test data set, our algorithm generated the color labeled clusters as shown in Figure 2(d).

The intent of the third synthetic example shown in Figure 3 is to explicitly demonstrate the added capabilities of our proposed distance measure over the more conventional chamfer and Hausdorff distance measures. The input data shown in Figure 3(a) contains two interlacing "U" shaped data sets. Specifically, the data sets are facing each other with each data set consisting of 3D parallel fiber tracts. By design, a tail of each data set is sandwiched between the wings of the other data set. As a result, the parallel fiber tracts near these tail regions are in close proximity to the parallel fiber tracts from the other data set. When only the Hausdorff (Figure 3(b)) or the chamfer (Figure 3(c)) distance is used as the similarity measure within a LLE manifold clustering algorithm, erroneous clustering results occur. However, our proposed similarity measure successfully clustered the input data as shown in Figure 3(d). This example illustrates the effectiveness of our proposed distance measure in capturing not only the local but also the global structure of the data set.

4.2. Fiber Traces from DT-MRI

The corpus callosum is a white matter structure located just ventral to the cortex that connects the left and right cerebral hemispheres to allow communication between the two halves of the brain. Subdividing the corpus callosum into anatomically defined portions is not well defined but of much importance, especially in study normal development and in understanding mental and neurodegenerative disorders. We apply our method in partitioning the fiber tracts of the corpus callosum into anatomical bundles. Figure 4 demonstrates the results of our fiber tract clustering

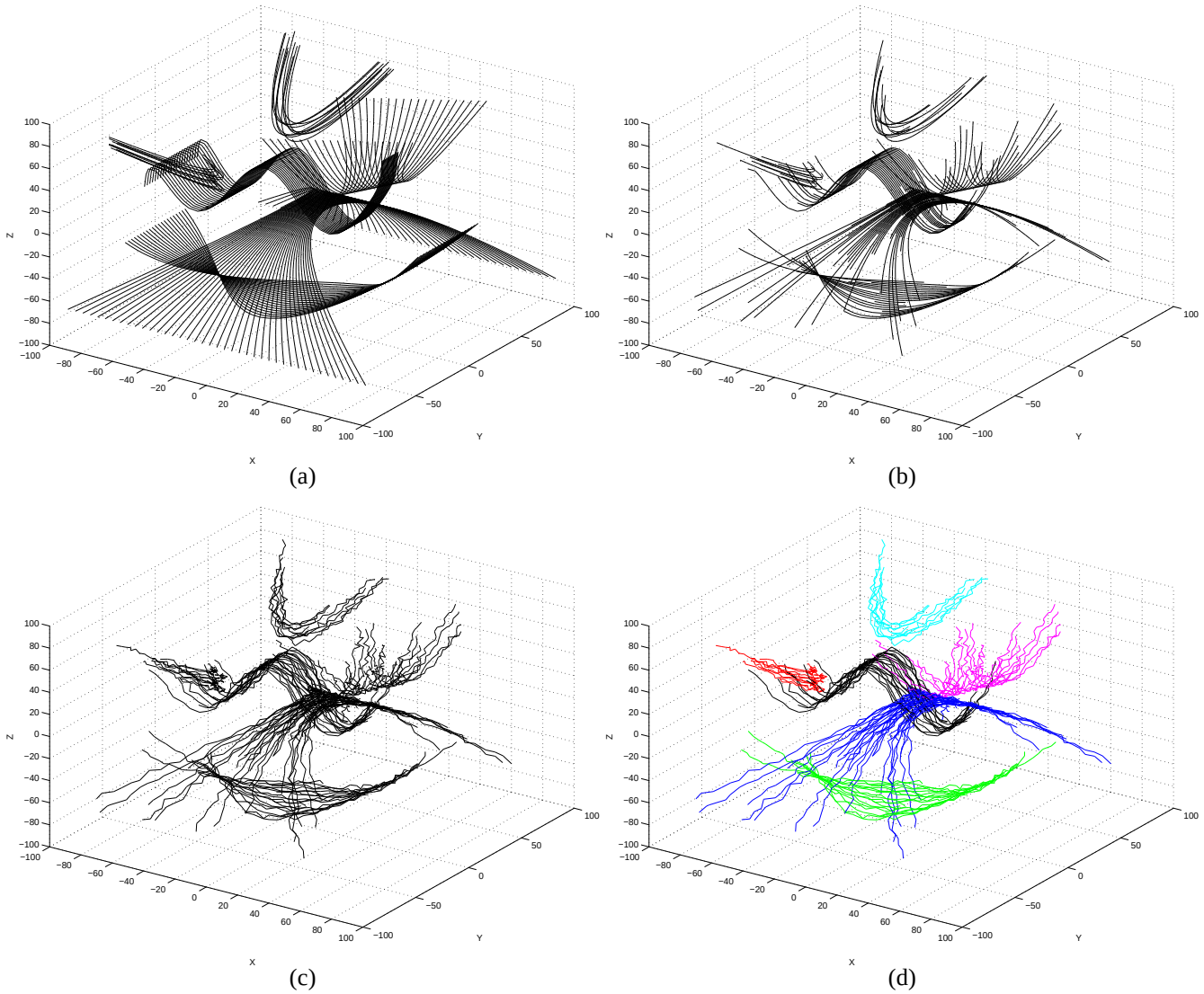


Figure 2. Synthetic fiber tract clustering example in a hostile environment. (a) Original data consisting of 126 synthetic fiber tracts each described by 50 data points. (b) Random removal of varying size segments to generate a data set with fiber tracts lengths varying between 4–50 data points. (c) Corruption of the data by additive Gaussian noise ($\mu = 0$ and $\sigma = 1$) to generate the input data set. (d) Correct partitioning of fiber tracts into 6 bundles with $K = 8$ and $d = 5$ in less than 23 seconds on a 3.1 GHz Intel Xenon processor.

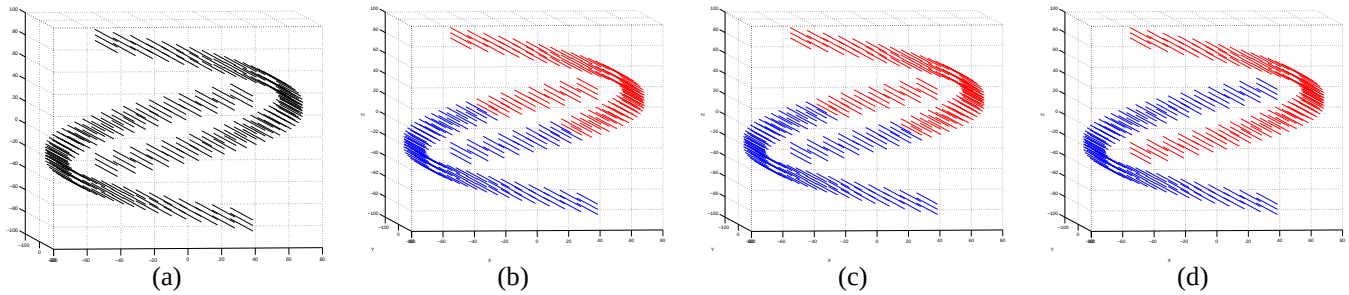


Figure 3. Synthetic example consisting of two interweaving "U" shaped data sets. (a) Input data consisting of 246 parallel fiber tracts. (b) Clustering result based on the Hausdorff distance. (c) Clustering result based on the chamfer distance. (d) Clustering result based on the proposed distance measure. ($K = 4$ and $d = 1$ for all these simulations.)

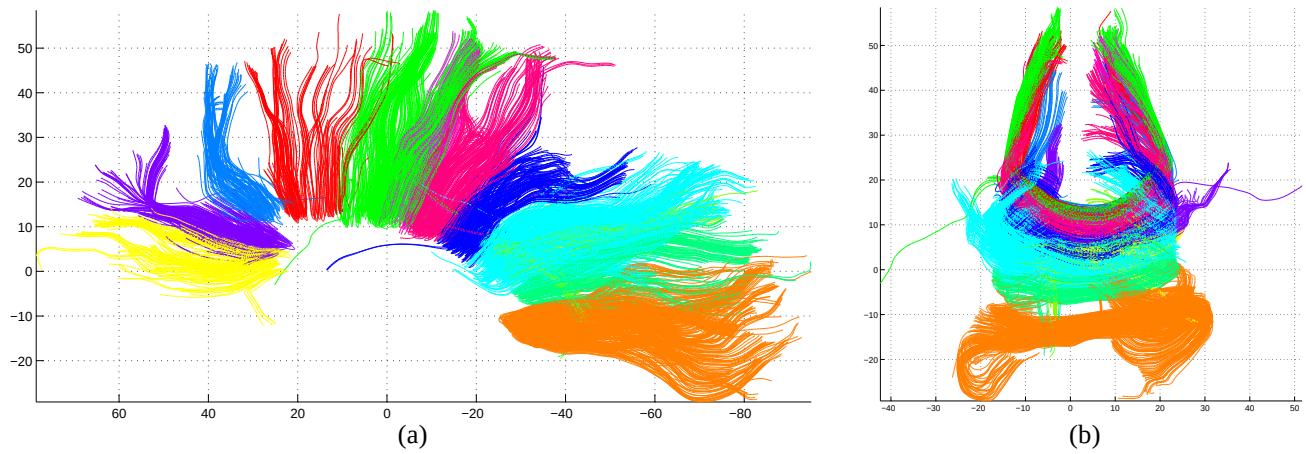


Figure 4. Clustering of 1665 corpus callosum fiber tracts obtained from diffusion tensor imaging with $K = 20$ and $d = 14$. (a) Sagittal view. (b) Coronal view.

algorithm in dividing a patient's 1665 corpus callosum fiber tracts into 12 clusters.

5. Conclusions and Future Work

We have outlined a manifold clustering algorithm utilizing Local Linear Embedding and a new similarity measure between fiber tracts built on the concept of dual rooted-graphs to yield a more robust and principled clustering algorithm that accounts for both the local and global geometry of the input data for effective partitioning. The preliminary results demonstrate improved visualization of the connectivity of the fiber tracts for clinical use. A natural extension of this work is to incorporate our proposed distance metric within other spectral clustering techniques. Much needed improvement is needed in speeding up the construction of the MSTs to reduce the calculation of the distances between the fiber tracts. We are actively investigating the clinical utility of our algorithm for disease understanding and treatment.

References

[1] H. Barrow, J. Tenenbaum, R. Bolles, and H. Wolf, Parametric correspondence and chamfer matching: Two new techniques for image matching, Int. Joint Conf. on AI, 1977. 2

[2] A. Brun, H. Knutsson, H.-J. Park, M. Shenton, and C.-F. Westin, Clustering fiber traces using normalized cuts, *MICCAI*, 2004. 2

[3] A. Brun, H. Park, H.-J. Knutsson, C.-F. Westin, Coloring of DT-MRI fiber traces using Laplacian eigenmaps, *EUROCAST*, 2003. 2

[4] R. Coifman and S. Lafon, Diffusion maps, *Applied and Computational Harmonic Analysis*, 2006. 3

[5] I. Corouge, S. Gouttard, and G. Gerig, Towards a shape model of white matter fiber bundles using diffusion tensor MRI, *Int. Symp. Biomed. Imag.*, 2004. 1

[6] Z. Ding, J. Gore, and A. Anderson, Classification and quantification of neuronal fiber pathways using diffusion tensor MRI, *Magnetic Resonance in Medicine*, 2003. 1

[7] S. Grikschat, J. Costa, A. Hero, and O. Michel, Dual rooted-diffusions for clustering and classification on manifolds, *ICASSP*, 2006. 2, 3

[8] D. Huttenlocher and W. Rucklidge, A multi-resolution technique for comparing images using the Hausdorff distance, *CVPR*, 1993. 2

[9] L. Jonasson, P. Hagmann, J.-P. Thiran, and V. Wedeen, Fiber tracts of high angular resolution diffusion MRI are easily segmented with spectral clustering, *ISMRM*, 2005. 2

[10] M. Maddah, A. Mewes, S. Haker, W. Grimson, and S. Warfield, Automated atlas-based clustering of white matter fiber tracts from DTMRI, *MICCAI*, 2005. 1

[11] S. Mori and P. Barker, Diffusion magnetic resonance imaging: Its principle and applications, *Anat. Record*, 1999. 1

[12] L. O'Donnell and C.-F. Westin, White matter tract clustering and correspondence in populations, *MICCAI*, 2005. 2

[13] L. Saul and S. Roweis, Think globally, fit locally: Unsupervised learning of low dimensional manifolds, *Journal of Machine Learning Research*, 2003. 3

[14] J. Shimony, A. Snyder, N. Lori, and T. Conturo, Automated fuzzy clustering of neuronal pathways in diffusion tensor tracking, *ISMRM*, 2005. 1

[15] S. Zhang, C. Demiralp, and D. Laidlaw, Visualizing diffusion tensor MR images using streamtubes and streamsurfaces, *IEEE Trans. Visual. and Comp. Graph.*, 9: 454–462, 2003. 1

[16] S. Zhang and D. Laidlaw, DTI fiber clustering and cross-subject cluster analysis, *ISMRM*, 2005. 1