2010 IEEE International Conference on Data Mining Workshops

Combining Time Series Similarity with Density-based Clustering to Identify Fiber Bundles in the Human Brain

Junming Shao*, Klaus Hahn†, Qinli Yang‡
, Christian Böhm*, Afra Wohlschläger§, Nicholas Myers§ and Claudia Plant*
*Institute for Computer Science, University of Munich, Munich, Germany
†Institute for Biomathematics & Biometrics, HMGU Helmholtz Center, Munich, Germany
‡School of Engineering, University of Edinburgh, Edinburgh, UK
§Department of Neuroradiology, Technische Universität München, Munich, Germany

***Abstract*—Understanding the connectome of the human brain is a major challenge in neuroscience. Discovering the wiring and the major cables of the brain is essential for a better understanding of brain function. Diffusion Tensor imaging (DTI) provides the potential way of exploring the organization of white matter fiber tracts in human subjects in a non-invasive way. However, it is a long way from the approximately one million voxels of a raw DT image to utilizable knowledge. After preprocessing including registration and motion correction, fiber tracking approaches extract thousands of fibers from diffusion weighted images. In this paper, we focus on the question how we can identify meaningful groups of fiber tracks which represent the major cables of the brain. We combine ideas from time series mining with density-based clustering to a novel framework for effective and efficient fiber clustering. We first introduce a novel fiber similarity measure based on dynamic time warping. This fiber warping measure successfully captures local similarity among fibers belonging to a common bundle but having different start and end points. A lower bound on this fiber warping measure speeds up computation. The result of fiber tracking often contains imperfect fibers and outliers. Therefore, we combine fiber warping with an outlier-robust density-based clustering algorithm. Extensive experiments on synthetic data and real data demonstrate the effectiveness and efficiency of our approach.**

***Keywords*-Diffusion Tensor Imaging, Dynamic Time Warping, Lower Bounding Distance, Density-based Fiber Clustering**

I. Introduction

Diffusion tensor imaging (DTI) provides a promising way to explore organization and integrity of white matter tracts in vivo, using water diffusion properties as a probe [1]. It captures the local diffusivity of water molecules within the tissue, and thus gives valuable insight about the course of fiber tracts since the water diffusion is along the direction of axons and restricted in the direction perpendicular to them. Potential pathways of fiber tracts in human brain thus can be reconstructed from diffusion weighted images (DWIs) via white matter tractography. This technique has attracted huge attention in neuroscience community to study anatomical connectivity [2], brain changes [3], etc.

However, after performing fiber tracking in human brain, we obtain thousands of tracks. The remaining problem in clinical research is how to classify these vast amount of fiber trajectories. To date, the most frequently used method is to select fibers based on expert knowledge, which is often referred to as *virtual dissection*. Experts first specify some regions of interest (ROIs) based on domain knowledge and then select all fibers that pass through these pre-defined ROIs [4]. This process tends to be inefficient especially since it is time consuming and limited by expert resources. Moreover, manual specification of ROIs in different patients may be biased. Therefore, a more promising approach is to cluster fiber tracks into fiber bundles which provide an overview on the structural organization of the fiber pathways. Grouping this large amount of data into more manageable clusters of tracks would be extremely useful for further applications. Therefore, the development of effective and efficient algorithms for white matter fiber tract segmentation in diffusion MRI is of significant interest for neuroscience community.

Currently, though several algorithms have been proposed for clustering trajectories into meaningful bundles, such as k-nearest neighbors [5], spectral clustering [6], [11], the following two problems are particularly vital but remain essentially unsolved:

1) Fiber Similarity Measure: An effective and efficient similarity measure for pair-wise fiber comparison is desired.
2) Outlier-robustness: Due to experimental limitations, like thermal noise or partial volume effects, the set of fibers produced by tractography contains also imperfect fibers or outliers. These fibers should not be clustered into any fiber bundle and need to be excluded.

In view of these issues, we propose a novel fiber similarity measure based on adapted dynamic time warping to calculate the pair-wise similarity distance between fibers. Further, a lower bounding technique for the similarity measure is proposed to save computation cost. Considering the noise in real data, we explore density-based clustering to group these fiber tracts. Imperfect fibers or outliers are eliminated during this clustering process.

The remainder of the paper is organized as follows. Re-

978-0-7695-4257-7/10 $26.00 © 2010 IEEE
DOI 10.1109/ICDMW.2010.15

lated work is reviewed and discussed in Section II. Section III and Section IV present our novel fiber similarity measure and density-based fiber clustering in detail. A series of experiments that we have performed and the relevant results are described in Section V. Finally, we conclude in Section VI.

II. Related work

In order to perform clustering, first a fiber similarity measure must be specified, which is then used as input to a clustering algorithm. Therefore, in this section, we will first give a brief survey on fiber similarity measures and then review recently proposed fiber clustering algorithms.

A. Fiber Similarity Measure

A fiber similarity measure is a function that computes the (dis)similarity between pairs of fibers. Two fibers are considered similar when they have comparable length, similar shape, and are separated by a small distance [5]. The early work by Brun et al. [7] assumes that two fiber tracts with similar end points should be considered as similar. Euclidean distance between the end points of fiber traces is then used to calculate the fiber similarity. However, the assumption is not reasonable in many cases since not all fiber bundles start and end in the same regions. It also ignores the information of most other points and the fiber pair-wise shape similarity. Ding et al. [5] propose a similarity measure by cutting each fiber into corresponding fiber segments and use the mean Euclidean distance between the segments to define piece-wise similarity. This similarity method is efficient but not effective since this measure also loses the point-by-point information. Several authors acknowledge that point-by-point correspondence of the trajectories should be used for accurate clustering and quantitative analysis. Zhang et al. [10] define the distance between two fibers as the average distance from any point on the shorter fiber to the closest point on the longer fiber, and only distances above a certain threshold contribute to this average. This similarity metric is thus not symmetrical and is affected by outliers or noise fibers. Corouge et al. [8] form point pairs by mapping each point of one fiber to the closest point on the other fiber. The resulting point pairs are then used to define the distance between fiber pairs. They define three distance measures: closest point distance, mean of closest point distance (MCP) and Hausdorff distance (HDD). The MCP and HDD measures are currently widely used for fiber similarity comparison. However, they are difficult to capture the fiber shape characteristics effectively, such as local shift or distortion. In order to overcome these problems, we present a novel fiber similarity measure based on dynamic time warping (DTW), to better represent the fiber similarity.

B. Fiber Clustering

Instead of grouping fibers by hand with expert knowledge, fiber clustering approaches take advantage of the similarity of the fiber paths to cluster these fibers by algorithms. Fiber clustering methods analyze a collection of white matter tracts in 3D and separate them into meaningful bundles, or clusters, that contain paths with similar shape and spatial position. These bundles are expected to contain fiber paths with similar anatomy and function. Corouge et al. [8] use a k-nearest neighbors method to calculate the similarity metric between paired fiber tracts defined in terms of the length ratio and the Euclidean distance. It propagates cluster labels from a fiber to a neighboring fiber, which assigns each unlabeled fiber to the cluster of its closest neighbor if the closest neighbor is below a threshold. A partition of the data with a specific number of clusters can be acquired by setting a threshold on the maximal accepted distance. This is similar to the algorithm employed by Ding et al. [5], which establish a corresponding segment to define the fiber similarity and then use $k-$nearest neighbors method to obtain fiber bundles. The k nearest-neighbors related method is very sensitive to the number of K and difficult to obtain the proper threshold without prior knowledge. Another popular fiber clustering method is spectral clustering [6], [11], which refers to a class of techniques which rely on the eigenstructure of a similarity matrix. This type of method first uses a spectral embedding technique to map the fibers to a new feature space and then uses traditional clustering method, such as K-Means to group the fibers in this new space. Brun et al. [6] use a spectral embedding technique called Laplacian eigenmaps to map the fibers to a Euclidean feature space and then use a Gaussian kernel to compare the fibers in this new space. Donnell et al. [11] decompose the fiber similarity matrix into the eigenvalues and eigenvectors. The top eigenvectors are used to represent each fiber and then, using a specific clustering called k-way normal cut, they obtain fiber bundles. Spectral clustering can detect arbitrarily shaped clusters but needs much memory and thus is not convenient to cluster large fiber sets. Furthermore, spectral clustering is sensitive to outliers and the user should specify the number of clusters, which is hard to know in advance. In this paper, we considers the fiber clustering problem from a density-based point of view, where fiber bundles are regarded as areas of high fiber density which are separated by areas of lower fiber density.

III. Fiber Warping

After deterministic tractography, a fiber is represented as an ordered set of points in space. The steps of the arc length, defined by two successive points of a fiber, are not necessarily identical, if e.g. numerical Runge Kutta methods with dynamical stepsizes are used for tracking. In addition, two fibers may have different lengths and consequently different numbers of points in space. To quantify similarity between two fibers, we adapt for this work the Dynamic Time Warping (DTW) method [12], [13].

In the following section, we will first briefly review DTW for time series and then extend it to a similarity measure

for space curves. Finally, we present a convenient lower bounding technique to save computational cost.

A. Dynamic Time Warping

Dynamic time warping (DTW) is a technique that looks for the optimal alignment of two time series. To achieve this goal, the time series are "warped" together non-linearly, by stretching or shrinking them along the time axes [14].

Suppose we have two time series X and Y, of lengths m and n respectively, where

$$X = (x_1, x_2, ..., x_i, ..., x_m) \tag{1}$$

$$Y = (y_1, y_2, ..., y_j, ..., y_n) \tag{2}$$

The objective is to optimize a warping path W:

$$W = (w_1, w_2, ..., w_k, ..., w_K) \tag{3}$$

where K is the length of W, with $max(m, n) < K < m + n - 1$. The k^{th} element of W is a pair of indices indicating a connection of time points in X and Y and is written as $w_k = (i, j)$, see Figure 1. A warping path follows the constraints [15] :

1) Boundary conditions: $w_1 = (1, 1)$ and $w_K = (m, n)$. This requires the warping path to start and finish in the first and last points of the series respectively;
2) Monotony: Given $w_k = (i, j)$, then $w_{k+1} = (i', j')$, with $i' - i >= 0$ and $j' - j >= 0$. This forces the points in W to be monotonically spaced in time.
3) Continuity: Given $w_k = (i, j)$, then $w_{k+1} = (i', j')$, with $i' - i <= 1$ and $j' - j <= 1$. This restricts the admissible steps in the warping path to adjacent points of the series.

There are many warping paths satisfying the above conditions. In order to find a best match between two time series, we look for that path which minimizes the cumulative distance between them. The distance dtw for this optimum path is defined as:

$$dtw(X, Y) = \min(\sum_{k=1}^{K} d(w_k)) \tag{4}$$

where $d(\cdot)$ is a distance function. We define it as

$$d(w_k) \equiv d(i, j) \equiv |x_i - y_j| \tag{5}$$

The optimum warping path for $dtw(X, Y)$ can be obtained through the dynamical programming approach [12]. It proceeds like follows: First, a m by n cost matrix D is constructed. A component $D(i, j)$ is defined recursively as sum of the distance $d(i, j)$ and the minimum of the cumulative distances in the adjacent elements :

$$D(i,j) = d(i,j) + min\{D(i-1, j-1), D(i-1, j), D(i, j-1)\} \tag{6}$$

After the entire cost matrix D is filled, starting from $D(1, 1)$, the minimum-distance warping path can be found in reverse order, starting from $D(m, n)$. For this purpose a greedy search is performed to evaluate cells to the left, down, and diagonally to the bottom-left. Whichever of these three adjacent cells has the minimum value is added to the beginning of the warping path found so far, and the search continues from that cell. The search stops if $D(1, 1)$ is reached. Figure 1 (a) shows an example of two time series with their cost matrix and a minimum-distance warping path between them. If the warping path passes through a cell $D(i, j)$ in the cost matrix, the i^{th} point in time series X is warped to the j^{th} point in time series Y. Since a single point may map to multiple points in the other time series, dynamic time warping can handle time series with different lengths. An illustration of the optimum warping path between two time series can be found in Figure 1 (b).

B. Fiber Similarity Measure with DTW

To calculate the similarity distance between fibers in space, we extend the one dimensional concept of dynamic time warping. Suppose $p_i(p_i^1, p_i^2, p_i^3)$ and $q_j(q_j^1, q_j^2, q_j^3)$ are points of the fibers P and Q, indexed along the arc length, where the three coordinates are given within the brackets. We define the distance between the two points as:

$$d(p_i, q_j) = |p_i^1 - q_j^1| + |p_i^2 - q_j^2| + |p_i^3 - q_j^3| \tag{7}$$

Using this distance function, we can treat the spatial fiber problem like a time series problem. The optimal warping path can be obtained through dynamical programming and is represented as $W = (w_1, ...w_k, ...w_K)$, where K is the length of the path and

$$d(w_k) \equiv d(p_i, q_j) \tag{8}$$

.

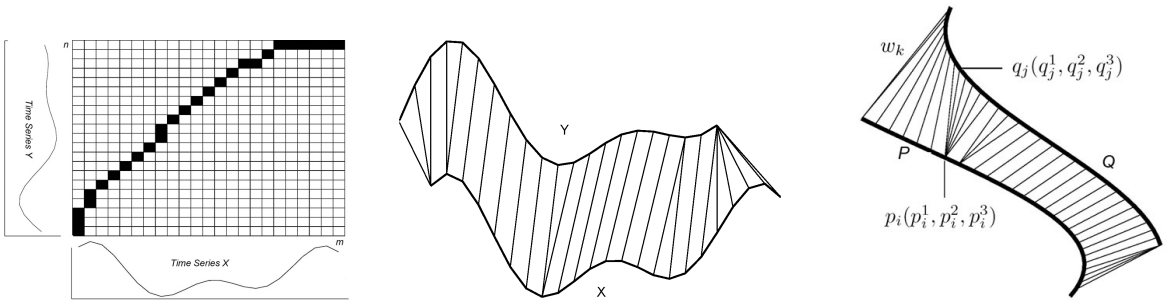
To reduce the effect of different lengths of fibers for similarity calculation, we define the similarity distance $DTW(P, Q)$ between the fibers P and Q as the averaged distance for the optimal warping path, see Figure 1 (c):

$$DTW(P, Q) = min(\frac{\sum_{k=1}^{K} d(w_k)}{K}) \tag{9}$$

C. Lower Bounding Distance

The time complexity of our fiber similarity measure DTW is $O(m \cdot n)$, which is demanding in terms of CPU time. To deal with this problem, we introduce an easily computed lower bounding distance LB. For two fibers we have $LB \leq DTW$, where an efficient LB should be a tight lower bound to DTW.

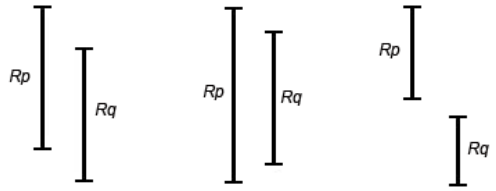
For two given spatial fibers $P(P^1, P^2, P^3)$ and $Q(Q^1, Q^2, Q^3)$, we rewrite them as three sequences of point pairs (P^1, Q^1), (P^2, Q^2) and (P^3, Q^3) respectively. Let us first consider the pair-wise sequence (P^1, Q^1). $Max(P^1)$



(a) Cost Matrix (b) Optimal warping path (time series) (c) Optimal warping path (fibers)

Figure 1. Cost Matrix (a) and optimal warping path for time series (b) and 3D fibers (c).

and $max(Q^1)$ denote the maximum values in P^1 and Q^1, respectively. $Min(P^1)$ and $min(Q^1)$ define the minimum values. A pair $(min(P^1), max(P^1))$ defines the range R_P of P^1. Without loss of generality, we assume $max(P^1) \geq max(Q^1)$. There are then three possible arrangements for the two ranges R_P and R_Q, see Figure 2.



(a) overlap (b) enclose (c) disjoint

Figure 2. Illustration of possible arrangements of R_P and R_Q.

The lower bounding distance between the two sequences is defined as follows [16]:

$$\mathsf{lb}(\mathsf{P}^1,\mathsf{Q}^1)=\begin{cases} \sum_{\mathsf{p}_i^1>\mathsf{max}(\mathsf{Q}^1)} |\mathsf{p}_i^1-\mathsf{max}(\mathsf{Q}^1)|+\sum_{\mathsf{q}_j^1<\mathsf{min}(\mathsf{P}^1)} |\mathsf{q}_j^1-\mathsf{min}(\mathsf{P}^1)| \\ \qquad\qquad \text{if } \mathsf{P}^1 \text{ and } \mathsf{Q}^1 \text{ overlap} \\ \sum_{\mathsf{p}_i^1>\mathsf{max}(\mathsf{Q}^1)} |\mathsf{p}_i^1-\mathsf{max}(\mathsf{Q}^1)|+\sum_{\mathsf{p}_i^1<\mathsf{min}(\mathsf{Q}^1)} |\mathsf{p}_i^1-\mathsf{min}(\mathsf{Q}^1)| \\ \qquad\qquad \text{if } \mathsf{P}^1 \text{ and } \mathsf{Q}^1 \text{ enclose} \\ \mathsf{max}(\sum_{i=1}^{|\mathsf{P}^1|} |\mathsf{p}_i^1-\mathsf{max}(\mathsf{Q}^1)|, \sum_{j=1}^{|\mathsf{Q}^1|} |\mathsf{q}_j^1-\mathsf{min}(\mathsf{P}^1)| \\ \qquad\qquad \text{if } \mathsf{P}^1 \text{ and } \mathsf{Q}^1 \text{ disjoint} \end{cases} \tag{10}$$

An analogous definition is used for the other pair-wise sequences (P^2, Q^2) and (P^3, Q^3). Finally, the lower bounding distance between fibers P and Q is defined as:

$$LB(P,Q) = \frac{lb(P^1,Q^1) + lb(P^2,Q^2) + lb(P^3,Q^3)}{m+n-1} \tag{11}$$

We prove in Appendix A the property : $LB(P,Q) \leq DTW(P,Q)$ for spatial fibers.

IV. Fiber clustering

A. Density-based clustering fiber tracts

After calculation of the fiber similarity, each fiber is regarded as a data object in metric space and a clustering approach can be applied to group the fibers. Since clustering is a common challenge in a large variety of applications, this problem has attracted much attention during the last decades, producing a vast number of research papers, books and surveys, eg. [17], [18], [19] to mention a few. One very interesting branch of research considers the clustering problem from a density-based point of view: Clusters are regarded as areas of high object density which are separated by areas of lower object density. Recently, some approaches to density-based clustering have been proposed. However, to the best of our knowledge none of them has been applied to the problem of fiber clustering so far. DBSCAN [18] is the most wide spread algorithm to density-based clustering and its definitions are intuitive in our context. The density-based clustering notion is formalized in DBSCAN using two parameters: ϵ specifying a range and $MinPts$ specifying a number of objects. The central notion of DBSCAN is the *core object*. A fiber is a core object of a density-based cluster if at least $MinPts$ fibers are in its ϵ-neighborhood. Formally this is captured by the following definitions:

Definition 1: (Core Object)
Let $\mathcal{D}$ be a set of n objects, $\epsilon \in \mathbb{R}^+$ and MinPts $\in \mathbb{N}^+$. An object $P \in \mathcal{D}$ is a core object, iff

$$|N_\epsilon(P)| \geq \text{MinPts, where } N_\epsilon(P) = \{Q \in \mathcal{D} : ||P-Q|| \leq \epsilon\}.$$

Two objects may be assigned to a common cluster. In density-based clustering this is formalized by the notions *direct density reachability, and density connectedness.*

Definition 2: (Direct Density Reachability)
Let $P, Q \in \mathcal{D}$. Q is called directly density reachable from P (in symbols: $P \lhd Q$) iff

1) P is a core object in $\mathcal{D}$, and
2) $Q \in N_\epsilon(P)$.

If P and Q are both core objects, then $P \lhd Q$ is equivalent with $P \rhd Q$. The density connectedness is the transitive and symmetric closure of the direct density reachability:

Definition 3: (Density Connectedness)
Two objects P and Q are called density connected (in symbols: $P \bowtie Q$) iff there is a sequence of core objects

```
algorithm Density-based Fiber Clustering
  Mark all objects as unprocessed.
  While(D contains unprocessed object) Loop
    Consider arbitrary unprocessed object P ∈ D
     //fiber range search with Lower-bounding distance
    N_ε(P) = rangeLBQuery(P,ε, D);
    If N_ε(P).size < MinPts
      assign P Noise
      Continue;
    Else
      assign new cluster-ID C
    EndIf

    For all elements Q ∈ N_ε(P)
      If Q is unprocessed
        mark element Q with cluster-ID C
        insert object Q into seed list S.
      EndIf
    EndFor

    While(S not ∅) Loop
      For all elements P' ∈ S
      N_ε(P') = rangeLBQuery(P',ε,D);
        If N_ε(P').size > MinPts
          For all elements Q' ∈ N_ε(P')
            If Q' is unprocessed or Noise
              If Q' is unprocessed
                insert object Q' into seed list S
              EndIf
            mark element Q' with cluster-ID C
            EndIf
          EndFor
        EndIf
      EndFor
    EndLoop
  EndLoop
```

Figure 3. Pseudocode of DBSCAN algorithm.

$(P_1, ..., P_m)$ of arbitrary length m such that

$$P \rhd P_1 \rhd ... \lhd P_m \lhd Q.$$

In density-based clustering, a cluster is defined as a maximal set of density connected objects:

Definition 4: (Density-based Cluster)
A subset $C \subseteq \mathcal{D}$ is called a cluster iff the following two conditions hold:

1) Density connectedness: $\forall P, Q \in C : P \bowtie Q$.
2) Maximality: $\forall P \in C, \forall Q \in \mathcal{D} \setminus C : \neg P \bowtie Q$.

The algorithm DBSCAN [18] implements the cluster notion of Definition 4 using a data structure called *seed list* S containing a set of seed objects for cluster expansion. More precisely, the algorithm proceeds as is indicated in Figure 3.

Like other clustering approaches, DBSCAN requires user to specify two parameters: ϵ and $MinPts$, which can be roughly estimated by visualizing the first "valley" of a sorted $k-dist$ graph [18].

V. Experimental result and analysis

In this section, we present a series of numerical experiments on synthetic and on real data to explore the efficiency and effectiveness of our approach. All algorithms are implemented in Java and the fiber visualization functions are programmed in MatLab. To compare clustering results for different approaches with a ground truth, an information-theoretic external cluster-validity measure [20] is used.

A. Fiber Data

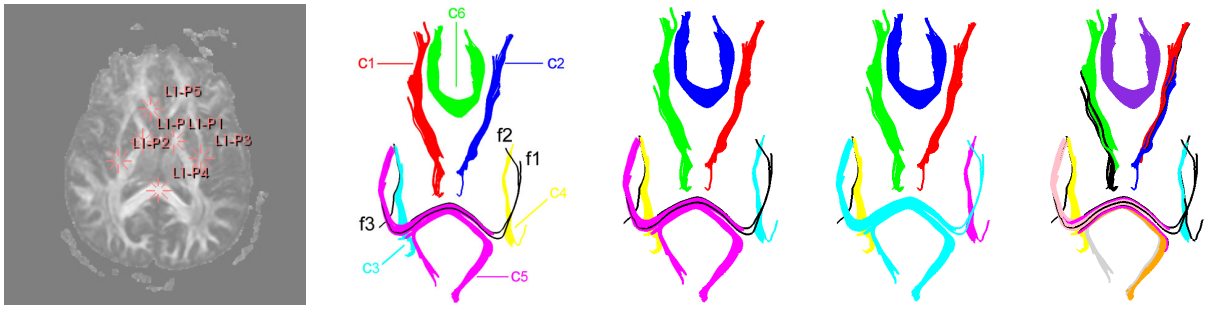
To evaluate our approach, realistic human brain data are taken from the open source software "Slicer3-3.4", folder "surgery case" (http://www.slicer.org/). This measurement comprises 55 non-collinear directions for the diffusion weighted images, with a $b-factor = 1000\ s/mm^2$, and 5 acquisitions for the reference, with a $b-factor = 0\ s/mm^2$. The brain volume contains $256 \times 256 \times 70$ voxels with size $1 \times 1 \times 2.6\ mm^3$. For fiber tracking the numerical Runge Kutta method (4-th order) is applied.

B. Experiments on Fiber Similarity Measure

To explore different similarity measures, a realistic set of fibers is obtained by specifying 6 fiducial seed regions manually (see Figure 4 (a) for the seed regions). They are located in the internal and external Capsules and in the Corpus Callosum. After fiber tractography, gold standard fiber clusters (ground truth) are created with the help of an experienced physician (see Figure 4 (b)). This gold standard includes 372 fibers and shows 6 anatomically meaningful fiber bundles assigned by C1-C6 with different colors and one group of noise or outliers (3 fibers, black color). These fibers $f1, f2, f3$ differ from their neighbours by shape and length.

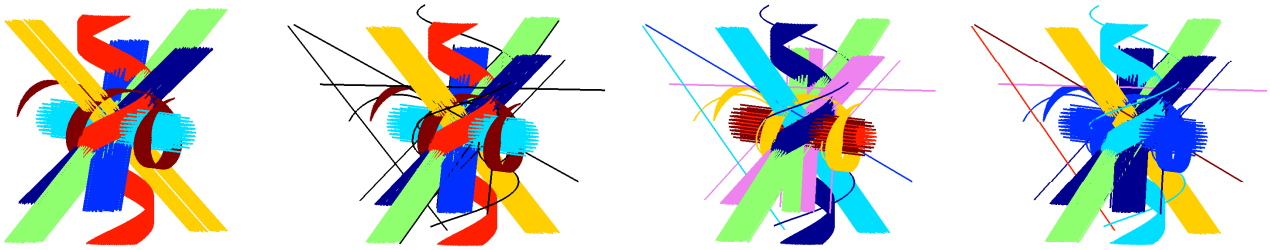
Based on these data, we compare our similarity measure (DTW) with two frequently used measures: Mean of closest point distance (MCP) and Hausdorff distance (HDD). For all three measures, our density-based clustering method is applied to achieve the result which is closest to the ground truth. It is clear that DTW and MCP separate the data into the 6 gold standard clusters of the ground truth, whereas HDD already segments the data for a global threshold into 10 clusters. For further comparison between DTW and MCP, we focus on the noise fibers indicated in Figure 4(b) in black color. Whereas MCP does not separate $f1, f2$ from the neighbors, DTW matches the ground truth. As MCP averages the minimum distances of point pairs from one fiber to another, it seems to "smooth out" the information in the data more than DTW.

Apart from the evaluation of effectiveness of a fiber similarity measure, the efficiency (computation cost) should also be considered. For this purpose we perform experiments on a range search (search the $\epsilon-$neighborhood fibers for one fixed fiber), which is the most time consuming step in fiber clustering. The involved similarity measures include: DTW with lower-bounding distance LB, DTW, MCP and HDD. The number of test fibers in the data set range from 1000 to 5000. Figure 5 presents the computational cost of range search for increasing numbers of fibers.



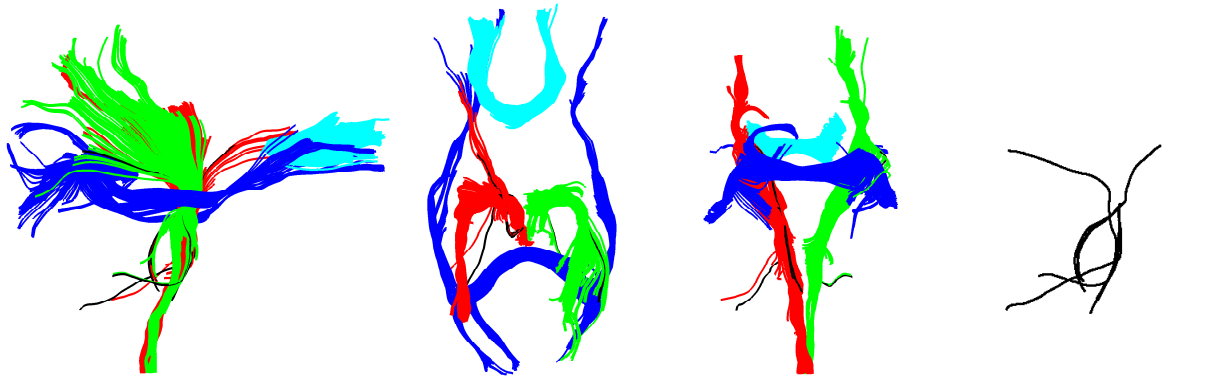
(a) Seed points (b) Gold standard (c) DTW (d) MCP (e) HDD

Figure 4. Experimental results of different fiber similarity measures. (a): Seed points, (b): Gold standard, (c-e): Fiber clustering results with different measures, where the same fiber bundles in identical coloring.



(a) Synthetic Data (b) Data with Outliers/DC (c) Spectral Clustering (d) Single Link

Figure 6. Different fiber clustering results based on synthetic data.



(a) Sagittal (b) Axial (c) Coronal (d) Sagittal (Noise)

Figure 7. The automatic fiber clustering result for Real Data 1 with parameters $MinPts = 6, \epsilon = 10$.

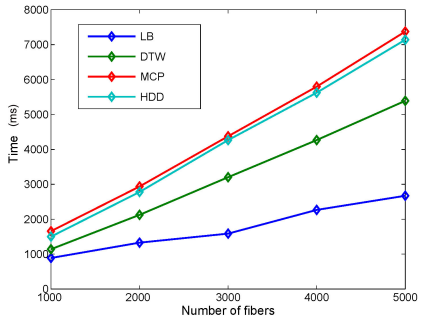
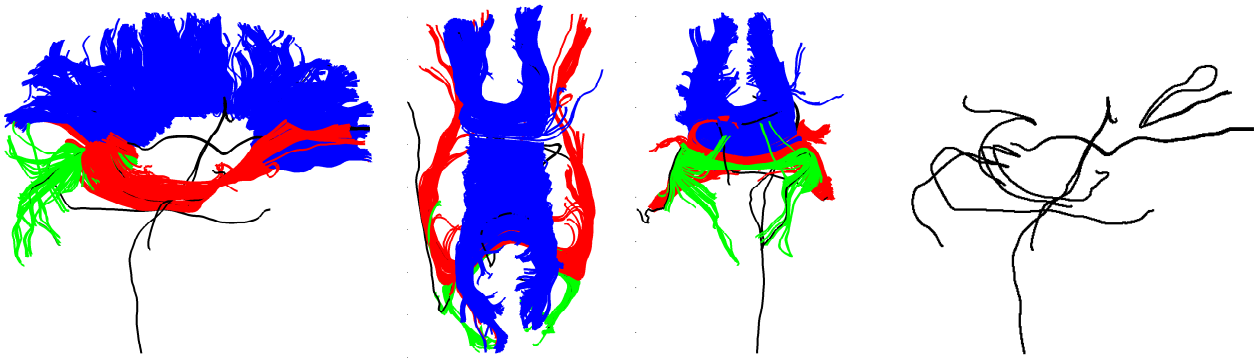


Figure 5. Comparison of efficiency with different fiber similarity measures.

C. Experiments on Fiber Clustering

In this section, we perform experiments on synthetic data as well as many real brain data to test our density-based fiber clustering scheme using our fiber similarity measure with lower-bounding technique. Other two fiber clustering approaches: Spectral clustering [19] and Hierarchical clustering (Single Link) [10] are also implemented in Java to compare with our fiber clustering approach.

1) Synthetic Data: The ground truth of our synthetic data in three dimensions includes five clusters of straight lines and two clusters of helices (see Figure 6 (a)), these clusters are composed of 410 individual fibers. Due to limitations of the experiment and of the tracking method, outliers may



(a) Sagittal (b) Axial (c) Coronal (d) Sagittal (Noise)

Figure 8. The automatic fiber clustering result for Corpus Callosum with parameters $MinPts = 6, \epsilon = 10$.

Table I
COMPARISON OF DIFFERENT FIBER CLUSTERING METHOD WITHOUT AND WITH OUTLIERS

Measures	Performance without and with Outliers		
	Conditional-entropy	*Code-length*	*Encoding-cost*
DC	0.0|0.0	0.304|0.358	**0.304**|**0.358**
SL	0.0|0.515	0.304|0.311	**0.304**|**0.826**
SC	0.0|0.775	0.304|0.232	**0.304**|**1.001**

appear in fiber data. To evaluate the three clustering methods with respect to such outliers, we add 8 outlier lines and 2 outlier helices to the synthetic data (Figure 6 (b)), creating a total of 420 synthetic fibers.

For all three clustering methods the ground truth presented in Figure 6 (a) is reproduced perfectly. The situation is different for Figure 6 (b). Only our approach can reproduce the ground truth, if the outliers are included, see (Figure 6 (b-d)). As we have the knowledge of the synthetic data (the true class label for each object), like the comparison of similarity measures, the information-theoretic external cluster-validity measure is applied to qualify the clustering results with different clustering methods. The Table I presents the quality of clustering results of data without and with outliers, which further indicates the effectiveness of our density-based fiber clustering.

2) Real Data: : In this section, we apply our approach on three real data sets with different clustering structures, to demonstrate the method's effectiveness.

Real Data 1: For Figure 7, we seeded 825 fibers from 4 different compact seed regions. From the mid-sagittal FA slice, one seed region is taken for occipital and temporal tracks and another one for prefrontal tracks. Two additional seed regions are taken in a symmetric way from a coronal FA slice to track projection fibers ending within the brain in the corona radiata [21]. Using the estimating parameters ($\epsilon = 10, MinPts = 6$) for clustering, Figure 7 (a-c), shows the 4 clusters are in agreement with the seed regions. Three fibers are viewed as outliers with black color (Figure 7 (d)).

Corpus Callosum: 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. From the mid-sagittal slice, 1100 seed points are taken from the corpus callosum via a FA map. Main part of the fibers connects cortical areas in approximate mirror-image sites. A smaller set is built of commissural trajectories to the temporal lobes [21]. With parameters ($\epsilon = 10, MinPts = 6$), three main fiber bundles and noisy fibers are successfully identified (Figure 8).

VI. CONCLUSION

In this paper, we have presented a novel framework for clustering white matter tracts. The technique combines a novel fiber similarity measure using adapted dynamic time warping and an outlier robust density-based clustering. The extensive experiments on synthetic as well as real data demonstrate the effectiveness and efficiency of our approach. To summarize, our fiber clustering method shows several desirable properties:

1) The adapted dynamic time warping is proposed to define fiber similarity, which allows to capture local similarity among fibers belonging to a common bundle but having different start and end points. It is more effective and efficient than MCP or HDD. In order to deal with the time complexity of fiber similarity, a lower-bounding technique is proposed for 3D fibers to speed up computational cost.
2) To explore meaningful fiber bundles in the human brain, we considers the fiber clustering problem from a density-based point of view. The number of clusters is not needed a priori and the DBSCAN parameters can be heuristically estimated by a $k - dist$ graph.
3) The density-based clustering approach can deal with noisy fibers which may be caused by limitations of imaging or by the fiber tracking technique. These noisy fibers are easy to be excluded from any cluster, which is of significant importance for further processing such as group tract-based analysis, etc.

APPENDIX A.
REMARK: LB IS A LOWER BOUND TO DTW FOR SPATIAL FIBERS

For two fibers $P = (p_1, ..., p_m)$ and $Q = (q_1, ..., q_n)$, for each dimension, we have $lb(p^d, q^d) < dtw(p^d, q^d), d = \{1, 2, 3\}$ [16]. we want to prove

$$DTW(P, Q) \geq LB(P, Q)$$

This is shown in the following.

$$\begin{aligned}
DTW(P,Q) &= \frac{\sum_{k=1}^{K}(|p_i^1 - q_j^1| + |p_i^2 - q_j^2| + |p_i^3 - q_j^3|))}{K} \\
&\geq \frac{\sum_{k=1}^{K}(|p_i^1 - q_j^1| + |p_i^2 - q_j^2| + |p_i^3 - q_j^3|))}{m+n-1} \\
&= \frac{\sum_{k=1}^{K}(|p_i^1 - q_j^1|)}{m+n-1} + \frac{\sum_{k=1}^{K}(|p_i^2 - q_j^2|)}{m+n-1} + \\
&\quad \frac{\sum_{k=1}^{K}(|p_i^3 - q_j^3|))}{m+n-1} \\
&= \frac{dtw(p^1, q^1) + dtw(p^2, q^2) + dtw(p^3, q^3)}{m+n-1} \\
&\geq \frac{lb(p^1, q^1) + lb(p^2, q^2) + lb(p^3, q^3)}{m+n-1} \\
&= LB(P, Q)
\end{aligned}$$

REFERENCES

[1] S. Mori, *Introduction to Diffusion Tensor Imaging*, New York: Elsevier, 2007.

[2] P. Hagmann, L. Cammoun, X. Gigandet, R. Meuli, C. J. Honey, V. J. Wedeen, and O. Sporns, "Mapping the structural core of the human cerebral cortex," P*LoS Biol*, Vol. 6, No. 7, e159, Jul. 2008.

[3] H. Huang, J. Zhang, S. Wakana, W. Zhang, T. Ren, L. J. Richards, P. Yarowsky, P.Donohue, E. Graham, P. C.M. v. Zijl, and S. Mori, "White and gray matter development in human fetal, newborn and pediatric brains," *Neuroimage*, vol. 33, pp. 27-38, 2006.

[4] M. Catani, R. J. Howard, S. Pajevic, and D. K. Jones, "Virtual in vivo interactive dissection of white matter fasciculi in the human brain," *NeuroImage*, vol. 17, pp. 77-94, 2002.

[5] Z. Ding, J. Gore, and A. Anderson, "Classification and quantification of neuronal fiber pathways using diffusion tensor MRI," *Magnetic Resonance in Medicine*, vol. 49, pp. 716C721, 2003.

[6] A. Brun, H. Park, H. J. Knutsson, and C. F. Westin, "Coloring of DT-MRI fiber traces using laplacian eigenmaps," In *proc. of EUROCAST 2003*, pp. 564-572, Feb. 2003.

[7] A. Brun, H.-J. Park, H. Knutsson, and C.-F. Westin, "Coloring of DTMRI fiber traces using Laplacian eigenmaps," *In Proc. the Ninth Int. Conf. on Computer Aided Systems Theory*, vol. 2809, pp. 564-572, Feb. 2003.

[8] I. Corouge, S. Gouttard, and G. Gerig, "Towards a shape model of white matter fiber bundles using diffusion tensor MRI," *IEEE Int. Symp. on Biomedical Imaging*, pp. 344-347, 2004.

[9] B. Moberts, A. Vilanova, and J. J. van Wijk, "Evaluation of Fiber Clustering Methods for diffusion tensor imaging," In *Proc. IEEE Visualization*, pp. 65-72, 2005.

[10] S. Zhang and D. H. Laidlaw, "DTI fiber clustering and cross-subject cluster analysis," In *proc. Int. Society for Magnetic Resonance in Medicine*, May 2005.

[11] L. ODonnell, M. Kubicki, M. E. Shenton, W. E. L. Grimson, and C.-F. Westin, "A method for clustering white matter fiber tracts," *American Journal of Neuroradiology*, vol. 27, no. 5, pp. 1032-1036, 2006.

[12] F. Itakura, "Minimum Prediction Residual Principle Applied to Speech Recognition," *IEEE Trans. Acoustics, Speech, and Signal Proc.*, vol. ASSP-23, pp. 52-72, 1975.

[13] H. Sakoe and S. Chiba, "Dynamic Programming Optimization for Spoken Word Recognition," *IEEE Trans. on Acoustics, Speech and Signal Proc.* vol. 26, pp. 623-625, 1978.

[14] S. Salvador and P. Chan, "FastDTW: Toward accurate dynamic time warping in linear time and space," 3rd Wkshp. on *Mining Temporal and Sequential Data, ACM KDD* '04, Seattle, Washington, Aug. 22-25, 2004.

[15] E. Keogh, "Exact Indexing of Dynamic Time Warping," In *proc. of the 28th Int. Conf. on Very Large Data Bases*, Hong Kong, pp. 406-417, Aug. 2002.

[16] B. Yi, H. Jagadish, and C. Faloutsos, "Efficient retrieval of similar time sequences under time warping," In *proc. of the 14th Int. Conf. on Data Engineering*, Orlando, FL, pp. 201-20, 1998.

[17] A. P. Dempster, N. M. Laird and D. B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society*, vol. 39, no. 1, pp. 1-31, 1977.

[18] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," *Proc. of the Second Int. Conf. on Knowledge Discovery and Data Mining*, pp. 226-231,1996.

[19] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," In *Advances in Neural Information Proc. Systems* vol. 14, pp. 849-856, 2001.

[20] B. Dom, "An information-theoretic external cluster-validity measure," *Technical Report RJ10219*, IBM, 2001.

[21] S. Mori, S. Wakana, L. M. Nagae-Poetscher, and P. C. M. Zijl, "MRI Atlas of Human White Matter", *Elsevier*, 2005.