Hierarchical Density-Based Clustering of White Matter Tracts in the Human Brain

Junming Shao, University of Munich Germany

Klaus Hahn, HMGU Helmholtz Center Munich, Germany

Qinli Yang, University of Edinburgh, UK

Afra Wohlschläeger, Technical University of Munich, Germany

Christian Boehm, University of Munich, Germany

Nicholas Myers, Technical University of Munich, Germany

Claudia Plant, Florida State University, USA

ABSTRACT

Diffusion tensor magnetic resonance imaging (DTI) provides a promising way of estimating the neural fiber pathways in the human brain non-invasively via white matter tractography. However, it is difficult to analyze the vast number of resulting tracts quantitatively. Automatic tract clustering would be useful for the neuroscience community, as it can contribute to accurate neurosurgical planning, tract-based analysis, or white matter atlas creation. In this paper, the authors propose a new framework for automatic white matter tract clustering using a hierarchical density-based approach. A novel fiber similarity measure based on dynamic time warping allows for an effective and efficient evaluation of fiber similarity. A lower bounding technique is used to further speed up the computation. Then the algorithm OPTICS is applied, to sort the data into a reachability plot, visualizing the clustering structure of the data. Interactive and automatic clustering algorithms are finally introduced to obtain the clusters. Extensive experiments on synthetic data and real data demonstrate the effectiveness and efficiency of our fiber similarity measure and show that the hierarchical density-based clustering method can group these tracts into meaningful bundles on multiple scales as well as eliminating noisy fibers.

Keywords: *DBSCAN, Density-Based Fiber Clustering, Diffusion Tensor Imaging, Dynamic Time Warping, Lower Bounding Distance, OPTICS*

INTRODUCTION

Diffusion Tensor imaging (DTI) can explore the organization and integrity of human white matter tracts in vivo, using water diffusion properties as a probe (Mori, 2007). It measures for every voxel the diffusivity of water molecules within the tissue, and thus gives valuable insight into the orientation of fiber tracts, since water diffusion is strongest along the direction of fibers

DOI: 10.4018/jkdb.2010100101

and restricted in the directions perpendicular to them. Potential pathways of fiber tracts in the white matter of the human brain can be reconstructed by deterministic and stochastic tractography based on the measured diffusion weighted images. These techniques have attracted attention in the study of anatomical connectivity (Hagmann et al., 2008), brain changes (Huang et al., 2006), and various pathologies related to white matter atrophy, such as schizophrenia (Park et al., 2004), Alzheimer's disease (Damoiseaux et al., 2009), or multiple sclerosis (Law & Grossman, 2005).

Performing fiber tracking in the human brain usually results in large sets of tracks. A problem in clinical research is how to segment and interpret this vast amount of information. A frequently used method is to select fiber groups of interest on the basis of expert knowledge by *virtual dissection*. Experts first specify some regions of interest (ROIs) and select then all fibers that pass through these pre-defined ROIs (Catani et al., 2002). This process tends to be inefficient, since the manual handling of ROIs is time-consuming and also limited by the availability of experts. Moreover, manual specification of ROIs may be biased in patient populations. Therefore, automatic clustering of fiber tracts into bundles of similar fibers is preferable for many applications. Two fibers are considered as similar if they have comparable length, similar shape and similar location (Ding, Gore, & Anderson, 2003). A number of approaches have been proposed to automatically cluster white matter tracts. However, before grouping fibers into clusters, it is necessary to specify a fiber similarity measure.

A fiber similarity measure is a function that quantifies the similarity between pairs of fibers. In the early work by Brun et al. (2003), it is assumed that two fiber tracts with similar end points should be considered as similar. The Euclidean distance between the starting and ending points of the two fibers is used to calculate fiber similarity. However, this assumption is not sufficient in some cases, as not all fibers in a bundle start and end in the same region. It also ignores the shape information contained in all points on the fibers. Ding, Gore, and Anderson (2003) propose a similarity measure by cutting each fiber into corresponding fiber segments and use then the mean Euclidean distance between the segments to define piece-wise similarity. This similarity method is efficient, but may lack effectivity since this measure loses the point-by-point information. Several authors acknowledge that the point-by-point correspondence of the trajectories should be included to define a convenient fiber similarity measure. For instance, in (Zhang, Demiralp, & Laidlaw, 2003; Zhang & Laidlaw, 2005) the distance is defined between two fibers as the average distance from any point on the shorter fiber to the closest point on the longer fiber, where only distances above a certain threshold contribute to this average. Another possibility is proposed by Klein et al. (2007). They introduce a grid-based similarity measure. A box around the fibers is partitioned into identical cells. Each point of a fiber is then assigned to the cells with weights. The similarity between a pair of fibers is calculated by the pairwise weights, summed over all cells including the fibers. The side length of the cell is an adjustable parameter that controls the scale of comparison. Corouge, Gouttard, and Gerig (2004) form point pairs, 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 similarity distances: closest point distance, mean of closest point distance (MCP) and Hausdorff distance (HDD).

Currently, MCP (Corouge, Gouttard, & Gerig, 2004; O'Donnell et al., 2006; O'Donnell & Westin, 2007; Moberts, Vilanova, & Wijk, 2005; Zhang & Laidlaw, 2005) and HDD (Corouge, Gouttard, & Gerig, 2004; Gerig, Gouttard, & Corouge, 2004; Moberts, Vilanova, & Wijk, 2005; Xia, Turken, Whitfield-Gabrieli, & Gabrieli, 2005) are widely used fiber similarity measures. Their formal definition is the following:

Suppose there are two fibers P and Q, then MCP is defined as:

$$d_{MCP}(P,Q) = avg\left(d_m(P,Q), d_m(Q,P)\right) \tag{1}$$

with:

$$d_m(P,Q) = avg_{p_i \in P}\{min_{q_j \in Q} \mid\mid p_i - q_j \mid\mid\}$$

While HDD is defined as:

$$d_{HDD}(P,Q) = avg\left(d_h(P,Q), d_h(Q,P)\right) \tag{2}$$

with:

$$d_h(P,Q) = \max_{p_i \in P}\{\min_{q_j \in Q} \mid\mid p_i - q_j \mid\mid\}$$

For the point-by-point fiber similarity measures, the computational cost or inefficiency is high as the time complexity is O($m \cdot n$), where m and n are the lengths of any two fibers respectively.

Having defined a convenient similarity measure, the fibers can be clustered. In the following we shortly review such algorithms.

Ding, Gore, and Anderson (2003) establish a k most similar-fibers algorithm to find automatically the different bundles. Based on a similarity measure for fiber segments, a cluster around fiber F is constructed by neighbouring fibers whose similarity to F is greatest. With similar clustering approach, Corouge, Gouttard, and Gerig (2004) use point pair distances for the similarity measure. Clusters of low cardinality are treated as outliers. The number of clusters is controlled by setting a threshold on the accepted maximum distance.

Another frequently used clustering method is spectral clustering (Brun, Park, Knutsson, & Westin, 2003; O'Donnell1 et al., 2006), which refers to techniques relying on the eigenstructure of the similarity matrix to partition fibers into disjoint clusters. Brun et al. (2003) use a spectral embedding technique called Laplacian eigenmaps to map the fibers to a Euclidean feature space and then they use a Gaussian kernel to compare the fibers in this new space, assigning similar fibers to similar colors. O'Donnell et al. (2006) decompose the fiber similarity matrix into their eigenvalues and eigenvectors. The top eigenvectors are used to represent the most important similarity information for each fiber, while removing noise. Then, a specific clustering called k-way normal cut is used to obtain fiber bundles. For this type of clustering the number of expected clusters must be defined a priori by the user. Klein et al. (2007) use an extension of multiple eigenvector clustering, a special variant of spectral clustering. This method includes an automatic detection of the number of clusters.

Zhang and Laidlaw (2005) use an agglomerative hierarchical clustering algorithm. This type of algorithm decomposes a fiber set into several levels of partitions, represented by a dendrogram. Starting with the clustering obtained by placing every object in a cluster, in every step the two closest clusters are merged until all objects are in one cluster. A drawback of hierarchical clustering is that it is hard to find the best partition of the dendrogram.

Maddah et al. (2005, 2006) incorporate anatomical expert knowledge into clustering. They use an atlas of fiber tracts, labeled by the number of their bundle. To build such an atlas, they start with a set of labeled ROIs specified by an expert and calculate the corresponding fiber bundles seeded from those ROIs. Affine registration is then used to map extracted fibers of some subject to the atlas. To make the comparison between fibers of the subject and of the atlas efficient, B-spline representations of the tracts are used. Similar subject and atlas fibers are labeled identically. O'Donnell and Westin (2007) also present an atlas based segmentation of fiber tracts. Their segmentation is performed in two steps. First, an atlas, labeled by experts, is learned from a population of subjects using spectral clustering. Based on this model, fibers of a novel subject are clustered by an extension of the spectral clustering solution stored in the atlas, using the Nystrom method. This segmentation across subjects enables testing of neuroscientific hypotheses with respect to group differences.

Though appreciable progress has been achieved in fiber clustering, the following problems are, in our opinion, still challenging and may justify a novel approach:

1) Fiber Similarity Measure: Effective point-by-point similarity measures for pair-wise fiber comparison are needed for clustering. These similarity measures should effectively capture similarities, but also efficiently, as big sets of fibers need to be analyzed.
2) Flexibility: The fiber clustering approach should be flexible enough to cluster fibers on multiple scales. Hierarchical cluster structures are prevalent in nature and can be found, e.g., for the fibers of the Corpus Callosum.
3) Transparency: A frequently used visualization of a hierarchical clustering result is the dendrogram. This graph lacks of transparency for large data sets and is hard to analyze. More informative visual presentations of a clustering structure would be useful.
4) Outlier-detection and -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. It is beneficial for further tract-based analysis to separate out those fibers during clustering to obtain compact clusters. Another aspect is the impact of such outliers on the results of some clustering algorithms. Algorithms which are robust to outliers are necessary.

In view of these issues, we propose in this paper a new fiber similarity measure which is adapted from dynamic time warping, to calculate a point-by-point similarity distance among fibers. Further, a lower bounding technique for this similarity measure is introduced, to reduce computational cost. Considering the different clustering substructures of fiber data and inherent outliers, we apply - to the best of our knowledge - for the first time hierarchical density-based clustering to group fiber tracts. The outlier-robustness of density-based clustering is demonstrated. The resulting clustering structure of the fiber data can be visualized in a transparent manner by a so-called reachability plot. From this plot, the clusters are subsequently calculated by interactive or automatic fiber clustering algorithms on multiple scales. Imperfect fibers or outliers are eliminated during the clustering process. The approach is tested on several synthetic as well as real fiber data sets.

The remainder of the paper is organized as follows: we first present the novel fiber similarity measure and the hierarchical density-based fiber clustering in detail. A series of clustering experiments on synthetic and real data and the relevant results are then described. A discussion of our results follows. Finally, we conclude our work.

FIBER SIMILARITY MEASURE

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 (Itakura, 1975; Sakoe & Chiba, 1978; Sankoff & Kruskal, 1983). DTW is introduced for the comparison of time series which are out of phase and is applied in fields as diverse as speech recognition (Sakoe & Chiba, 1978), bioinformatics (Aach & Church, 2001) or data mining (Keogh & Pazzani, 2001).

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.

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 (Salvador & Chan, 2004).

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

$$X = (x_1, x_2, \ldots, x_i, \ldots, x_m) \quad (3)$$

$$Y = (y_1, y_2, \ldots, y_i, \ldots, y_m) \quad (4)$$

The objective is to optimize a warping path W:

$$W = (w_1, w_2, \ldots, w_k, \ldots, w_K) \quad (5)$$

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 (Keogh, 2002):

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:

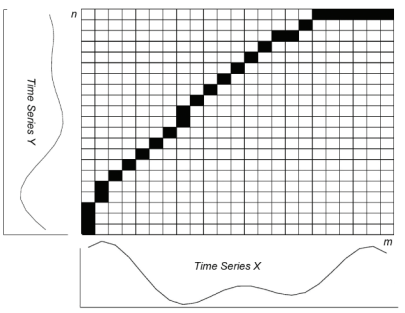
$$dtw(X,Y) = \min\left(\sum_{k=1}^{K} d(w_k)\right) \quad (6)$$

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

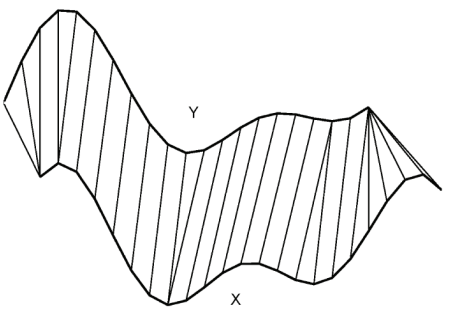
$$d(w_k) \equiv d(i, j) \equiv |x_i - y_j| \quad (7)$$

The optimum warping path for *dtw*(*X*, *Y*) can be obtained through the dynamical programming approach (Itakura, 1975). 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:

Figure 1. Cost matrix (a) and optimal warping path for time series (b)



(a) Cost Matrix



(b) Optimal warping path (time series)

$$D(i,j)=d(i,j)+\min\{D(i-1,j-1),D(i-1,j),D(i,j-1)\} \quad (8)$$

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. The warping path is W = {(1, 1), (1, 2), (1, 3), (2, 4), (2, 5), (3, 6), (4, 7), (5, 8), (6,9), (7, 10), (8, 11), (8, 12), (9, 13), (10, 14), (11, 15), (12, 16), (13, 17), (14, 18), (15, 19), (16, 19), (17, 20), (18, 21), (19, 21), (20, 21), (21, 21), (22, 21), (23, 21)}. 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. However, we should note that the dynamic time warping distance function doesn't satisfy triangle inequality and thus not a metric. An illustration of the optimum warping path between two time series can be found in Figure 1 (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:

$$d(p_i, q_j) = | p_i^1 - q_j^1 | + | p_i^2 - q_j^2 | + | p_i^3 - q_j^3 | \quad (9)$$

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;

$$d(w_k) \equiv d(p_i, q_j) \quad (10)$$

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:

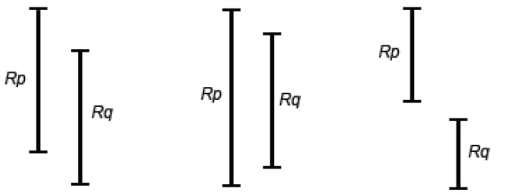
$$DTW(P,Q) = min(\frac{\sum_{k=1}^{K} d(w_k)}{K}) \quad (11)$$

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). $Max(P^1)$ and $max(Q^1)$ denote the maximum values in P^1 and Q^1, respectively. Min(P1) and min(Q1) 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 $max(P^1)$ $max(Q^1)$. There are then three possible arrangements for the two ranges R_p and R_Q, see Figure 2.

Figure 2. Illustration of possible arrangements of R_P and R_Q



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

The lower bounding distance between the two sequences is defined as follows (Yi, Jagadish, & Faloutsos, 1998):

$$lb(P^1,Q^1) = \begin{cases} \sum_{p_i^1 > max(Q^1)} |p_i^1 - max(Q^1)| + \sum_{q_j^1 < min(P^1)} |q_j^1 - min(P^1)| \\ \quad if\ P^1\ and\ Q^1\ overlap \\ \sum_{p_i^1 > max(Q^1)} |p_i^1 - max(Q^1)| + \sum_{p_i^1 < min(Q^1)} |p_i^1 - min(Q^1)| \\ \quad if\ P^1\ and\ Q^1\ enclose \\ max\left(\sum_{i=1}^{|P^1|} |p_i^1 - max(Q^1)|, \sum_{j=1}^{|Q^1|} |q_j^1 - min(P^1)| \right) \\ \quad if\ P^1\ and\ Q^1\ disjoint \end{cases} \quad (12)$$

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:

$$LB(P,Q) = \frac{lb(P^1,Q^1) + lb(P^2,Q^2) + lb(P^3,Q^3)}{m+n-1} \quad (13)$$

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

This concept is especially useful for density-based clustering, see the following Section. For clustering, it is necessary to perform a so-called range search. That means, it is necessary to query the data set for the most similar fibers with respect to P, i.e. for fibers Q with *DTW*(P, Q) below a given threshold ε. First, *LB*(P, Q) is computed for these fibers. *DTW*(P, Q) must then only be computed for those fibers with $LB(P, Q) < \varepsilon$. The Pseudocode of ε-range search with lower bounding distance is provided in Figure 3.

HIERARCHICAL DENSITY-BASED FIBER CLUSTERING

After calculation of the fiber similarities, each fiber is regarded as an object in metric space, and a clustering approach can be applied to group fibers. Clustering has attracted much attention during the last decades, which led to the publication of a vast number of research papers, books, and surveys (e.g., Ankerst, Breunig, Kriegel, & Sander, 1999; Dempster, Laird, & Rubin, 1977; Ester, Kriegel, Sander, & Xu, 1996; Ng, Jordan, &Weiss, 2001; Jain & Dubes, 1988). One 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 low density. This notion has several attractive benefits: In contrast to other methods, the users do not have to specify the number of clusters they want to find. Density-based clustering is able to detect clusters of arbitrary shape. It is robust to noise and outliers. Finally, density-based clustering methods are not restricted to vector data, but are applicable to general metric spaces. A number of approaches have recently been proposed, including DBSCAN (Ester, Kriegel, Sander, & Xu, 1996), DBCLASD (Xu, Ester, Kriegel,

Figure 3. Pseudocode of ε -range query using lower bounding distance

```
algorithm rangeLBQuery (P, ϵ, 𝒟)
  Neigbors = {};
  For each object Q ∈ 𝒟
      LB_dist = LB(Q, P);
      if(LB_dist < ϵ)
          DTW_dist = DTW(Q, P);
          if(DTW_dist < ϵ)
              Neigbors.add(Q)
          EndIf
      EndIf
  EndFor
Return Neigbors;
```

& Sander, 1998), DENCLUE (Hinneburg & Keim, 1998), and OPTICS (Ankerst, Breunig, Kriegel, & Sander, 1999).

DENSITY-BASED CLUSTERING

The idea behind density-based clustering is, that the object density in neighbourhood of each cluster object P is sufficiently high. The basic algorithm for density-based clustering is DBSCAN (Ester, Kriegel, Sander, & Xu, 1996). In DBSCAN a flat clustering of the data is computed. A hierarchical structure, containing subclusters embedded in bigger clusters, cannot be detected. On the other hand, DBSCAN does not require the user to know the number of clusters a priori and can find clusters of arbitrary shape. Clustering with DBSCAN relies on two parameters: *MinPts*, specifying a minimum number of objects in a neighbourhood, and ε , specifying the radius of the neighbourhood.

These two parameters determine a density threshold for clustering. DBSCAN discriminates three types of objects. An object is called a core object, if it is in the interior of a cluster. In this case at least *MinPts* objects are in its ε -neighbourhood. A border object is in the neighbourhood of a core object, but contains fewer objects than *MinPts* in its eps-neighbourhood. A noise object is neither a core nor a border object. Any two core objects, which are close enough-within the distance ε - are put into the same cluster. Any border object that is close enough to a core object is put into the same cluster as the core object. Noise objects are stored elsewhere (cf., Ester, Kriegel, Sander, & Xu, 1996) for a formal description of DBSCAN), see (Shao et al., 2010) for a first application of DBSCAN to the problem of fiber clustering.

OPTICS (Ankerst, Breunig, Kriegel, & Sander, 1999) is a hierarchical density-based clustering algorithm, which emerges from DBSCAN. In contrast to DBSCAN, OPTICS does not assign cluster membership, but orders objects of a dataset D into a hierarchical cluster structure. This structure contains clustering information, which could also be achieved by all possible DBSCAN clusterings with respect to distances ε ' that are smaller than the generating distance ε . To describe the hierarchical cluster structure, each object is characterized by two measures, its core-distance and its reachability-distance. Formally this is captured by the following definitions:

Definition 1: (Core-distance of an object P) Let $P \in D$, $MinPts \in N$, $\varepsilon \in R$, $N_\varepsilon(P)$ be the ε -neighbourhood of P, $MinPts_{dist}(P)$ be the distance from P to its *MinPts*-nearest neighbour. The core-distance of object P is defined as:

$$CoDist(P) = \begin{cases} UNDEFINED, \mid N_{\varepsilon}(P) \mid < MinPts \\ \\ MinPts_{dist}(P), otherwise. \end{cases}$$

The core-distance of an object P measures its local density. It is defined as the *MinPts*-nearest neighbour distance of P if P is a core object. Otherwise it is UNDEFINED.

Definition 2: (Reachability-distance of O with respect to object P) Let P, O ∈ D, *MinPts* ∈ N, $\varepsilon \in R$, $N_{\varepsilon}(P)$ be the ε-neighbourhood of P, the reachability distance of object O w.r.t P is defined as:

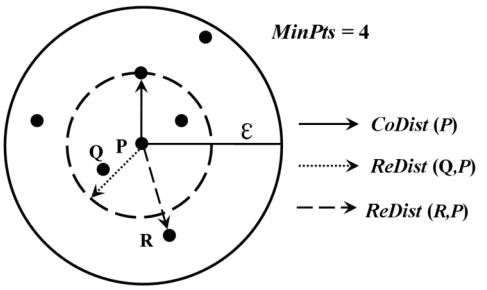
$$ReDist(O,P) = \begin{cases} UNDEFINED, \mid N_{\varepsilon}(P) \mid < MinPts \\ \\ max(CoDist(P), dist(O,P)), otherwise \end{cases}$$

Figure 4 illustrates the notions of core-distance and reachability-distance. The reachability-distance of Q w.r.t. P is the core-distance of P while the reachability-distance of R w.r.t. P is the distance between P and R.

The result of the algorithm OPTICS is an ordering of the data set reflecting the hierarchical cluster structure. More precisely, during the run of the algorithm, OPTICS computes for each object the core distance and a suitable reachability distance and writes this information to an output file. OPTICS starts with an arbitrary unprocessed object O, sets its reachability distance to UNDEFINED and determines its core distance. The object O is now processed and written with its core and reachability distance to the output. If O is a core object, all objects in the ε-neighbourhood of O are retrieved and inserted into a data structure called *seed list*. The seed list is a priority queue and the objects in the seed list are sorted according to the minimum reachability distance with respect to any of the objects processed before. As next object, OPTICS always selects the top object of the seed list, or in case the seed list is empty, an arbitrary unprocessed object from the data set. Consider the case that the start object O is a core object and object P is the top object in the seed list which has minimum reachability distance w.r.t. O. Object P is processed now, which implies that P is written with its core distance and reachability distance to the output. If P is a core object, all objects in the ε neighbourhood are added to the seed list. If necessary, the seed list is updated: Some objects already in the seed list may have a smaller reachability distance from P than from O. The algorithm terminates as soon as all objects have been processed. In this ordering, for every object the smallest reachability distance w.r.t. the preceding objects is determined.

Finally, the structure of the data ordering can be visualized in a reachability plot. This reachability plot is a 2D plot showing the objects' position index on the *x*-axis and

Figure 4. Illustration of core-distance and reachability-distance. The radius of the solid circle indicates the generating distance ε while the radius of the dashed circle is the MinPts-nearest neighbour distance of P (core-distance).



the objects' reachability-distance on the *y*-axis. The reachability plot provides a visual representation of the cluster structure of the data from the density-based point of view and enables an intuitive way to find clusters. Valleys in this plot indicate clusters, the deeper the valley, the denser the cluster. Clusters can be obtained by cutting the reachability plot with a horizontal line. An example of a reachability plot for a 2D data set is presented in Figure 5. The reachablity plot provides information about the general number of clusters, the densities of different clusters, and the hierarchical structure of the data. Another visual exploration paradigm which can simplify navigation and clustering of complex fiber structures was introduced by (Jianu, Demiralp, & Laidlaw, 2009). They use dendrograms from hierarchical clustering together with a two dimensional embedding of tract similarity based on a spring model following Hook's law.

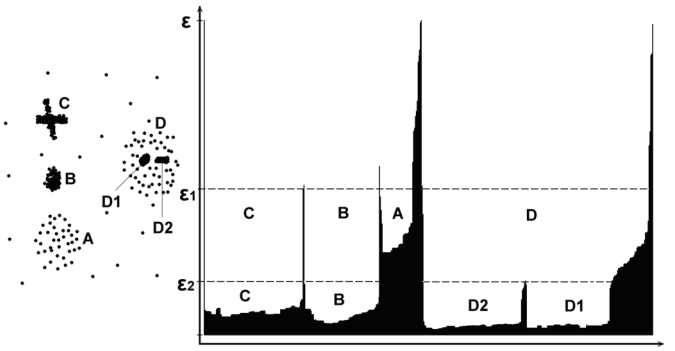
Like DBSCAN, OPTICS requires the two parameters: *MinPts* and ε. For DBSCAN it is not trivial to find the optimal parameters for clustering (Ester, Kriegel, Sander, & Xu, 1996). OPTICS, however, is rather insensitive to the input parameters (Ankerst, Breunig, Kriegel, & Sander, 1999). For all experiments in this paper which cover very different data structures *MinPts*=10 and ε=30 have been used. See the section of Discussion for the parameter selection.

FIBER CLUSTER EXTRACTION

The simplest way to obtain clusters in a reachability plot is to select a suitable parameter $0<\varepsilon_i<\varepsilon$ and to cut the reachability plot with the corresponding horizontal line. Intuitively, the valleys below the line indicate the clusters. Also their left borders are part of the clusters, if they are not noise. Noise or outliers of clusters are defined by the condition that their core- and reachability distances are larger than ε_i. See for an illustration Figure 5. With parameter ε_1, we obtain four clusters A, B, C, and D. If the line is moved down to ε_2, we obtain four new clusters B, C, D1, and D2, and some data which are interpreted as outliers (e.g., the cluster A and the group on the right side of D1). This demonstrates that a global threshold may not be sufficient, to extract all meaningful clusters (e.g., A with ε_2), since clusters can have varying densities. Therefore, two improved methods are proposed for fiber clustering.

Interactive fiber cluster extraction. To extract fiber bundles with different densities, it

Figure 5. Illustration of a reachability plot. Left: 2D sample data set, where different capital letters indicate hierarchical clusters with different density, shape and size. Right: the reachability plot of the sample data set, the different valleys of this plot indicate the clusterse.



is desirable to use a series of localized parameters. We have implemented a novel software to enable such interactive fiber clustering. The parameters can be selected from a screen-graph of the reachability plot in several steps.

1) For a first segmentation of the fibers, we select a global ε_1 (first level). This may create several large clusters ($C_1^1,\ldots,C_i^1,\ldots,C_n^1$) and one cluster O for noise, where the core- and reachability distances are above ε_1. Quantification of noise follows the definition proposed for DBSCAN.
2) For each cluster C_i^l found at level *l*, there may exist several valleys inside. To obtain those smaller subclusters, we go to the next level $l=l+1$, selecting the parameter ε_i^l only within C_i^{l-1} and proceed like in step 1). This creates new subclusters and in some cases noise.
3) To find clusters at higher levels, we repeat step 2) until the fibers are appropriately segmented.

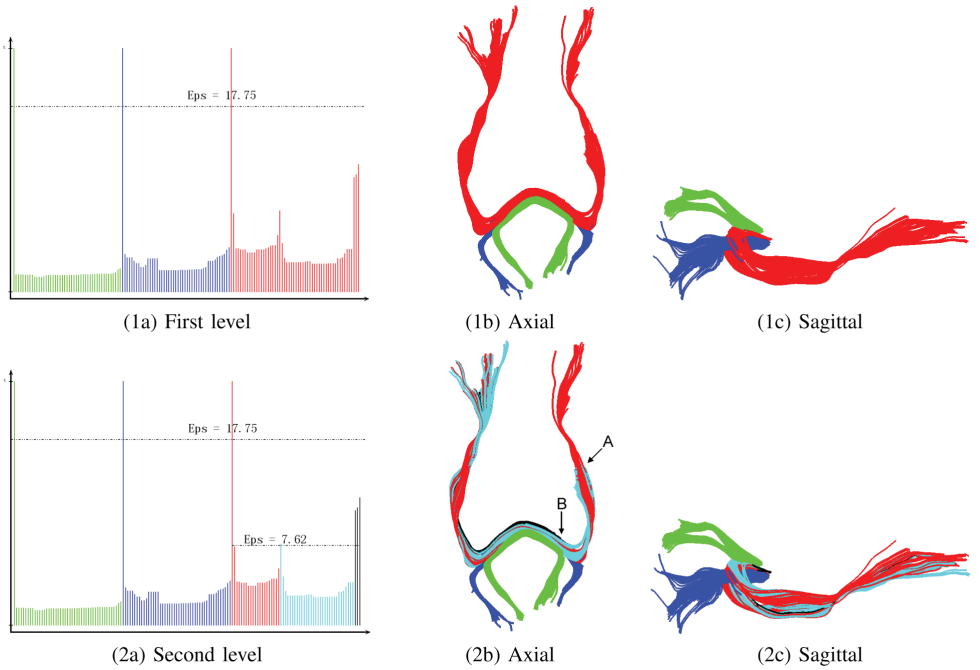
To select the parameter values ε on the screen-graph, users only need to "double-click" one position (x,y) in the reachability plot. The y-coordinate of the position indicates the ε, the x-coordinate is used to indicate its range. The first range is automatically global, while further ranges are localized to the cluster chosen. Figure 6 illustrates this approach on a real data set. Here, 250 fibers are seeded in the Splenium of Corpus Callosum (see the Experimental Section for details of the data and of the tractography). For the first segmentation into three clusters ε_1=17.75 is chosen (see Figure 6 (1a-1c) for the segmented reachability plot and for the fiber clusters in identical coloring). At the second level, we use ε_2=7.62, to separate the two obvious valleys within the red cluster. See Figure 6 (2a-2c) for the resulting clusters and for the three noise fibers. The arrows in Figure 6 2b indicate the endpoints of the cluster colored in cyan (A) and of the noise cluster (B), colored black. This fast and easy to use software provides users with a flexible method to quantify a visually meaningful clustering of a reachability plot.

Automatic fiber cluster extraction. The interactive method is useful to get a segmentation of data sets with a simple cluster structure, which can be inspected visually. For more complex data sets, however, an automatic segmentation is needed. To achieve this goal, we adapted the Tree Clustering algorithm (Sander, Qin, Lu, Niu, & Kovarsky, 2003), which extracts a hierarchical clustering from a reachability plot and creates a cluster tree. The basic idea in this algorithm is to identify significant local maxima separating the valleys in the reachability plot and to use these points for cluster extraction. A local maximum is regarded as significant if its height is well above the level of the valleys to its right and left side.

To determine the clusters between the local maxima, two parameters are introduced: *MinSize*, specifying the minimum number of fibers in a cluster or the minimum scale of segmentation, and *Ratio*, which determines if the local maximum is sufficiently above the neighbouring valleys to use it as a split point between two clusters.

We give a short description of our adaption of the Tree Clustering algorithm, which can now also segment highly irregular reachability plots. First, we collect all points P whose reachability distances (RD) are local maxima of the right and left neighbourhoods that enclose at least *MinSize* points. This set of points P is then sorted in descending order according to their RD P.r. Clusters are determined from this list by recursively removing the point P with largest RD, possibly separating clusters, until the list is empty. P is regarded as split point SP of two neighbouring clusters, if the median values of the RDs in both valleys are significantly lower

Figure 6. Illustration of the interactive clustering approach on a real dataset. (1a): interactive fiber cluster extraction at the first level based on the reachability plot. (1b-1c): The corresponding segmented fiber bundles in Axial and Sagittal perspectives. Different colors indicate different fiber bundles. (2a): Fiber cluster extraction at the second level. (2b-2c): The segmented fiber bundles at the second level, where the fiber bundles with red color are further split into two new fiber bundles: one with cyan color ending at A and another with the red color. Three fibers with black color ending at B are interpreted as noise.



than *SP.r* (*median*/*SP.r* < *Ratio* < 1). Noise points N within a cluster are detected by the condition *N.r* > *SP.r* and *CoDist*(N) > *SP.r*. Clusters without noise are listed in a tree graph, and noise is collected separately. There are two places in the tree where new nodes can be added: they can become children of the current node, or children of the parent node, replacing the current node. If the RDs of the current and parent split points are close, the newly created nodes are attached to the parent node instead of the current node itself.

Figure 7 illustrates the application of this method to the same data set which was used for interactive clustering. The reachability plot (Figure 7(a)) and the cluster tree (Fig. 7(b)) indicate the same fiber clustering and noise, which is presented in Figure 6. *MinSize*=10 and *Ratio*=0.7 are applied. See section of Discussion for the parameter setting.

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 fiber clustering approach. All algorithms are implemented in Java and the fiber visualization functions are programmed in MatLab. The calculations have been performed on a workstation with 2.4 GHz CPU and 2.0 GB RAM. We first present a comparison of fiber similarity measures and evaluate then different clustering

algorithms with respect to outliers. Finally, an application of hierarchical density-based fiber clustering to special real fiber sets is presented.

CLUSTER VALIDATION MEASURE

To compare clustering results for different approaches with a ground truth, an information-theoretic external cluster-validity measure (Dom, 2001) is used. We give a short description of this measure: A clustering of a set of fibers can be described by a set of cluster labels, mapping every fiber to a cluster. The evaluation scheme measures how useful the calculated cluster labels are as predictors of the ground truth cluster labels. In contrast to measures like the Rand Index (Halkidi, Batistakis, & Vazirgiannis, 2001) or the Cluster Purity (Wang, Wirth, & Wang, 2007), different numbers of clusters can be compared. The clustering quality measure is defined as the entire encoding cost, Q(C, K), which is the sum of the empirical conditional entropy H(C|K) and of the code length for the number of clusters CL(C|K):

$$Q(C, K) = H(C|K) + CL(C|K) \quad (14)$$

Where C is the set of ground truth labels and K equals the set of calculated labels.

$$H(C \mid K) = -\sum_{c=1}^{|C|}\sum_{k=1}^{|K|}\frac{h(c,k)}{n} log \frac{h(c,k)}{h(k)}$$

$$CL(C \mid K) = \frac{1}{n}\sum_{k=1}^{|K|} log\left(h(k) + \mid C \mid -1\right)$$

where h(c, k) is the number of fibers labeled within class C with label c and within class K with label $h(k) \equiv \sum_c h(c,k)$. $|K|$ and $|C|$ is the number of calculated clusters and the number of ground truth clusters respectively. The smaller the value of the entire encoding cost, the better is the quality of the clustering results.

FIBER DATA

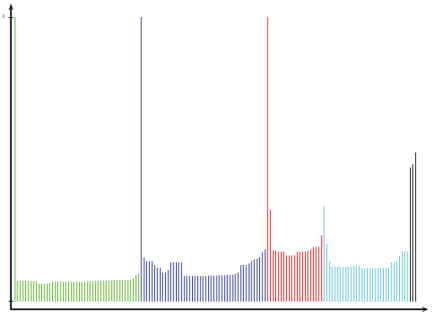
To evaluate our approach, realistic human brain data are taken from the open source software "Slicer3-3.4", folder "surgery case" (http://www.slicer.org/). This measurement comprises 55 non-collinear directions for the diffusion weighted images, with a *b-factor*=1000 s/mm^2, and 5 acquisitions for the reference, with a *b-factor* =0 s/mm^2. The brain volume contains 256×256×70 voxels with size 1 × 1×2.6 mm^3. For fiber tracking the numerical Runge Kutta method (4-th order) is applied. The seed points are given in two ways: First, small fiducial seed regions are defined by specifying their central points within the map of Fractional Anisotropy (FA); second, ROI seeding is defined by all seed points within a larger constrained volume. The deterministic tracks for fiber clustering are started from the seed points, following the principal diffusion in both directions. They are stopped for too low FA and for too high curvature of the tracks.

EXPERIMENTS ON FIBER SIMILARITY MEASURE

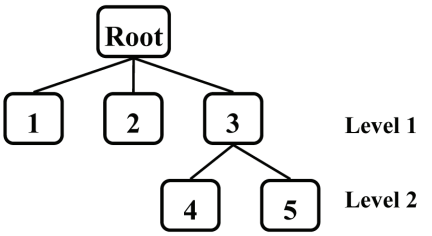
To explore different similarity measures, a realistic set of fibers is obtained by specifying 6 fiducial seed regions manually (see Figure 8 (a) for the seed regions). They are located in the internal and external Capsules and in the Corpus Callosum. After fiber tractography, gold standard fiber clusters (ground truth) are created with the help of an experienced physician (See Figure 8 (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 theses 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 the reachability plots are calcu-

Figure 7. Illustration of automatic fiber clustering. (a): Automatic fiber clustering based on the reachability plot with ratio= 0.7, minSize = 10 . (b): The cluster structure of the data set.



(a) Reachablity Plot



(b) Fiber Cluster Structure

lated and presented in Figure 8 (1a-1c). The interactive clustering method is applied to achieve the result which is closest to the ground truth. For this purpose a global threshold is sufficient, see Figure 8 (2a-2c).

It is clear from the reachability plots that DTW and MCP separate the data into the 6 gold standard clusters (6 big valleys) of the ground truth, whereas HDD already segments the data for a global threshold into 10 valleys. The reachability plot of DTW shows more local maxima than that of MCP, indicating more information about possible cluster separations or noise. For further comparison between DTW and MCP, we focus on the noise fibers indicated in Figure 8 (1a-1b) in black color. Whereas MCP does not isolate *f1*, *f2* from the neighbours, 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 ε – neighbourhood 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 9 presents the computational cost of range search for increasing numbers of fibers.

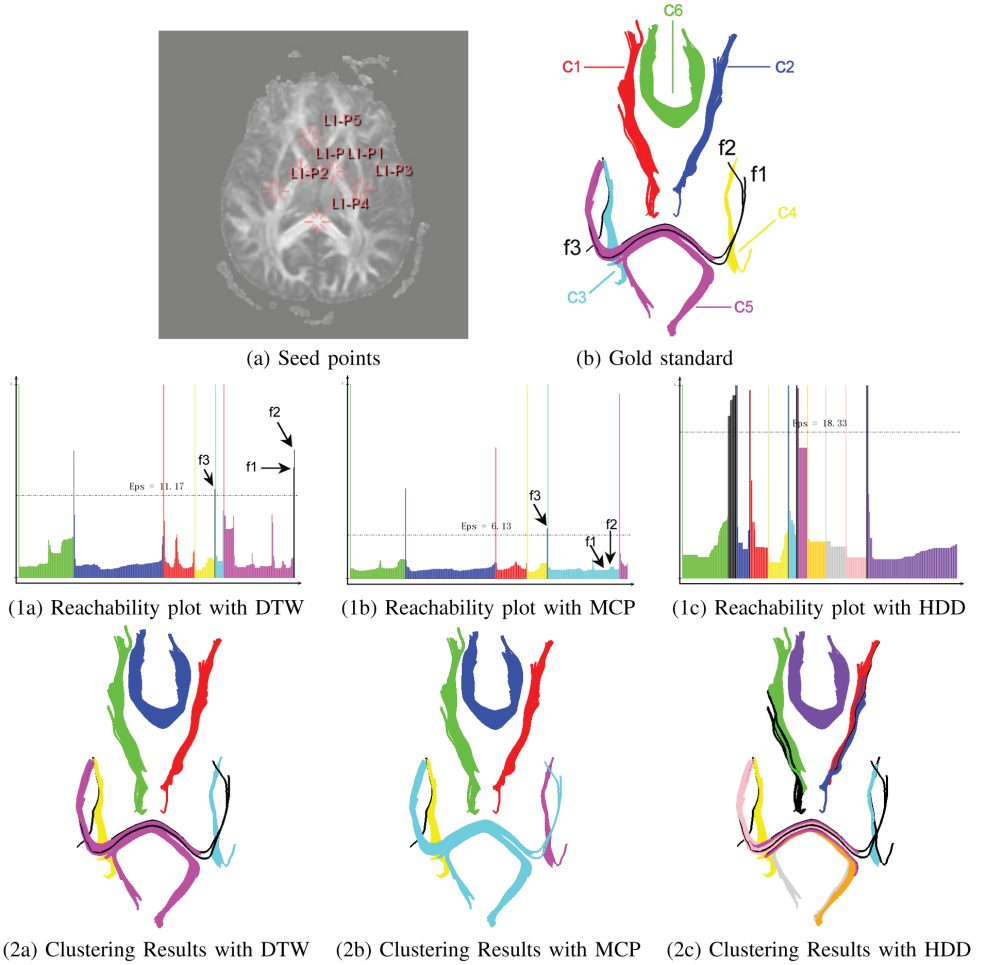
In Figure 9 we see that the efficiency of DTW outperforms that of MCP and HDD. LB further improves the efficiency of DTW. The larger the size of the data, the more the advantage of the LB technique becomes apparent. For example, LB costs about 60% and 35% of the computation time needed for MCP, when they are applied to 1000 and 5000 fibers respectively. Compared with DTW, LB improves efficiency by about 50% for 5000 fibers.

EXPERIMENTS ON FIBER CLUSTERING

In this section, we perform experiments on synthetic data as well as on real brain data using DTW with LB. The clustering method OPTICS is first compared to two different approaches: Spectral clustering (SC) (Ng, Jordan & Weiss, 2001), and Hierarchical clustering (Single Link, SL) (Zhang, Demiralp, & Laidlaw, 2003).

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 10 (a)), these clusters are composed of 410 individual fibers. To test the sensitivity to variations within the fibers we add gaussian noise to

Figure 8. Experimental results of different fiber similarity measures. (a): Seed points, (b): the gold standard, (1a-1c): reachablity plots for DTW, MCP and HDD, (2a-2c): the corresponding fibers in identical coloring.



(a) Seed points (b) Gold standard

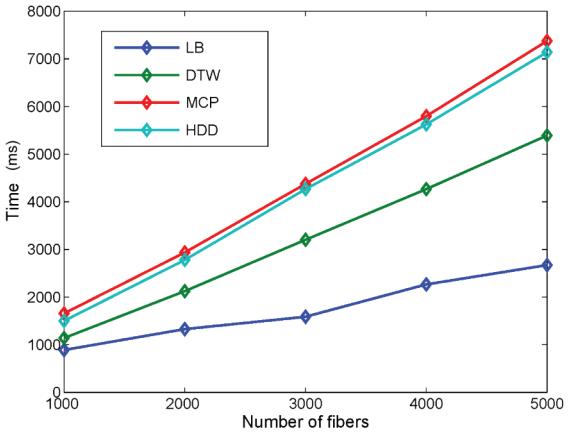
(1a) Reachability plot with DTW (1b) Reachability plot with MCP (1c) Reachability plot with HDD

(2a) Clustering Results with DTW (2b) Clustering Results with MCP (2c) Clustering Results with HDD

the fibers. For every point $(x(t), y(t), z(t))$ of a synthetic fiber, we modify the coordinates by $(x(t)+\varepsilon(t), y(t)+\varepsilon(t), z(t)+\varepsilon(t))$, where $\varepsilon(t)$ is Gaussian noise (Figure 10 (b)). Moreover, due to limitations of the experiment and of the tracking method, outliers may 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 10 (c)), creating a total of 420 synthetic fibers.

For all three clustering methods the ground truth presented in Figure 10 (a-b) is reproduced perfectly. The situation is different for Figure 10 (c). Only OPTICS can reproduce the ground truth, if the outliers are included, see (Figure 10 (d-f)). For SC the user has to specify the desired number of clusters (8 clusters are applied in (Figure 10 (e)). For SL either the perfect split of the dendrogram has to be chosen or, like in SC, the number of clusters must be given (8 clusters are chosen). For the perfect split some information theoretic methods, like

Figure 9. Comparison of efficiency with different fiber similarity measures



Minimal Description Length (Hansen & Yu, 2001), can be used.

Figure 11(a-b) shows the dendrograms and reachability plots for the synthetic fibers without (410 fibers) and with the 10 additional noisy fibers. Only the top 30 nodes in the dendrogram are visualized: the total number of layers is 419. We see that the dendrogram is appreciably complicated by adding a few outliers. The reachability plots in both cases, however, are robust to the outliers and already indicate visually the 7 clusters and the outliers.

These results are quantified in Table 1, where the conditional-entropy, the code-length and the Encoding-cost are given for the synthetic data without and with outliers. The strong impact of only few outliers on the clustering results is apparent for SL and SC.

Applications of hierarchical density-based clustering to real data. In this section, we apply OPTICS and automatic clustering on two real data sets with very different clustering structure, to demonstrate the method's flexibility.

Real Data 1: For Figure 12, we seeded 1768 fibers from 5 fiducial seed regions, chosen according to (Mori, Wakana, Nagae-Poetscher, & Zijl, 2005). Six big fiber groups are thus created, comprising two branches of the Corticospinal Tracts (red, magenta), part of the middle Cerebellar Peduncle (cyan), two parts of the Cingulum (yellow, blue) and a smaller section of the Corpus Callosum (green). Figure 12 (a) shows the reachability plot and its colored segmentation, which corresponds to the cluster structure in Figure12 (b). A low *Ratio*=0.2 and *MinSize*=30 are used to achieve a coarse clustering. Finer details of the structures within the clusters should not be detected. Figures 12 (1a-1c) present the result in different perspectives. The six clusters are well detected, but some noise is found as well. Noise is presented in Figure 12 (1d).

Real Data 2: Figure 13 presents a segmentation of the Corpus Callosum. Starting from the mid-sagittal slice of an FA-map within the Corpus Callosum, 1100 fiber tracts are calculated by ROI seeding of the slicer data. The main part of the fibers connects cortical areas in approximate mirror-image sites. A smaller subset is built of commissural trajectories to the Temporal Lobes (Mori, Wakana, Nagae-Poetscher, & Zijl, 2005). Automatic fiber clustering is applied to the reachability plot of these data with *Ratio*=0.7 and *MinSize*=30 to detect also subclusters within any hierarchy. The final

reachability plot with colored segmentation and the cluster tree are given in Figure 13 (a-b). Automatic Tree Clustering creates three levels, as illustrated in Figure 13 (1a-1c). Noise fibers are presented in black color. At the first level, only three big clusters are detected. This flat clustering result can also be achieved by an application of DBSCAN, see (Shao et al., 2010). Now, the clustering hierarchy is detected. The second level splits up cluster number three (blue color) into four clusters. At the third level, a finer segmentation into nine clusters of similar fibers is performed. Noise is given separately in Figures 13 (2a-2c). The final compact clustering without noise is given in Figures 13 (3a-3c) in three perspectives.

DISCUSSION

We present in this paper a novel framework for clustering of white matter tracts. First, we introduce a novel similarity measure DTW between two fibers which is based on dynamical time warping. This method, developed for time series which are out of phase, is adapted to comparisons between space curves. Then, a convenient lower bound is introduced to speed up the calculation of this measure. Second, the robust hierarchical density-based clustering method OPTICS is applied. To the best of our knowledge, this is the first application of OPTICS to the fiber clustering problem. The outcome is a reachabiltiy plot, which gives a transparent visualization of the clustering structure of the fibers. The clusters or valleys in the reachability plot are then determined by a novel interactive method and by a new modification of an automatic method, called Tree Clustering. Noise or outlier fibers are detected in a local fashion.

To compare DTW with the standard measures, mean of closest point distance (MCP) and the Hausdorff distance (HDD), a gold standard is introduced. We show that the reachability plot for DTW perfectly reproduces the gold standard, whereas MCP fails to detect the outliers and HDD cannot even detect the main clusters. This indicates that DTW more effectively maps the information about clustering into the reachability plot than MCP and HDD do. Also efficiency is highest for DTW, which can even be improved by the use of a lower bound technique.

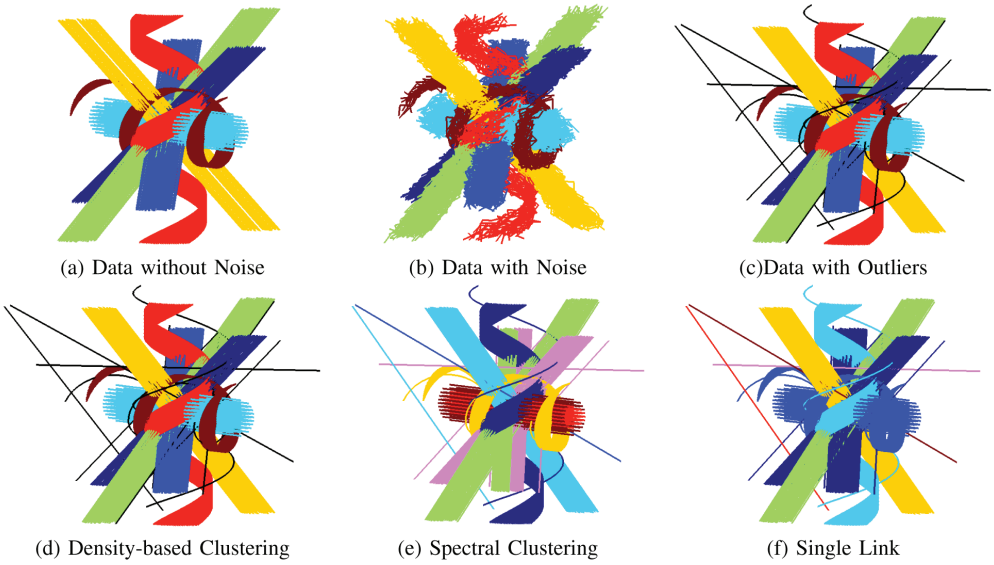
We further compare the clustering method Spectral Clustering and the hierarchical method Single Link with OPTICS. Synthetic examples show no advantage of OPTICS for regular data, but more robustness if outliers are included.

The result of OPTICS is a reachability plot visualizing the clustering structure of the data. This plot must be further analyzed to quantify the location of the clusters. For this purpose we introduce an easy to use interactive software and demonstrate its utility on a realistic fiber set seeded in the Splenium of Corpus Callosum. The main bundles are reasonably detected, also noise fibers are found. The same results are achieved with an automatic method. This method is an adaption of Tree Clustering to the irregular reachability plots based on DTW. To make Tree Clustering more robust, we use the median to calculate the significance of the split points and eliminate noise fibers at every recursive clustering step. Noise or outliers are identified by a criterion defined in density based clustering (Ankerst, Breunig, Kriegel, & Sander, 1999). Our method detects noise in a local fashion, as at

Table 1. Comparison of different fiber clustering without and with outliers

Measures	Performance without and with Outliers		
	Conditional-entropy	*Code-length*	*Encoding-cost*
OPTICS	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**

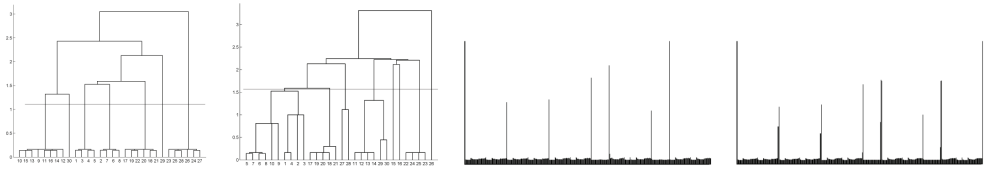
Figure 10. Different fiber clustering results based on synthetic data



(a) Data without Noise (b) Data with Noise (c)Data with Outliers

(d) Density-based Clustering (e) Spectral Clustering (f) Single Link

Figure 11. Dendrograms and reachability Plots for the synthetic data



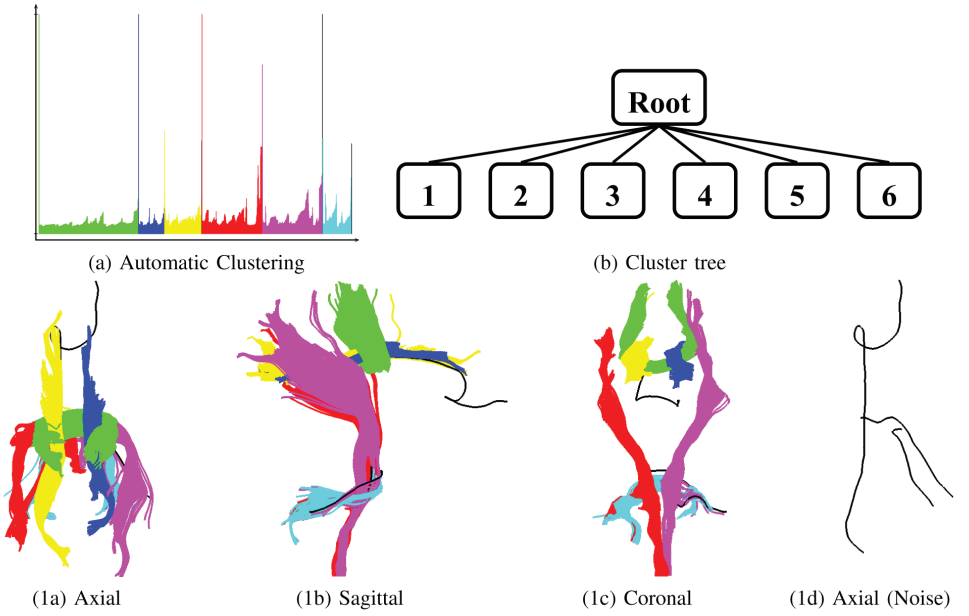
Data without outliers Data with outliers Data without outliers Data with outliers

(a) Dendrogram Tree (b) Reachability Plot

every level of the cluster hierarchy noise can be detected. A fiber in a cluster is defined as noise, if its core and reachability distances are above the reachability distance of the corresponding split point. Other concepts of local outliers or noise in density based clustering can be found in (Breunig, Kriegel, Ng, & Sander, 2000). A comparative exploration of the different noise concepts is beyond the scope of this work.

Finally two very different realistic data sets are segmented, to demonstrate the flexibility of our method. For data set 1, including partly the Cingulum (left, right branch), the Corpus Callsoum, the Corticospinal Tracts (left, right branch) and the Cerebellar Peduncle, a coarse clustering is performed. This is achieved by special parameters in Tree Clustering. For *Ratio*=0.2 the finer details of any hierarchy in the clusters are suppressed and *MinSize*=30 prevented too small clusters. The cluster tree shows just one level of a flat clustering. Ignoring any hierarchy, the fiber clusters fit to a segmentation into the 6 big groups of the ground truth. For data set 2 the Corpus Callosum is segmented, a hierarchy should be detected. For this purpose the parameters *Ratio*=0.7 and *MinSize*=30 are applied and produced three levels in the hierarchical cluster tree. The clustering fits well to

Figure 12. Automatic fiber clustering on the real data 1. (a): Automatic clustering of the fibers with parameter ratio=0.2, MinSize=30 based on the reachablity plot. (b): The fiber cluster structure of the data, where the data are split into 6 fiber bundles; (1a-1c): Fiber clustering results, the different color indicates various fiber bundles and black color means outliers. (1d): The outliers in axial perspective.



the anatomical expectation, looking for groups with similar fibers. At the end 11 clusters are detected.

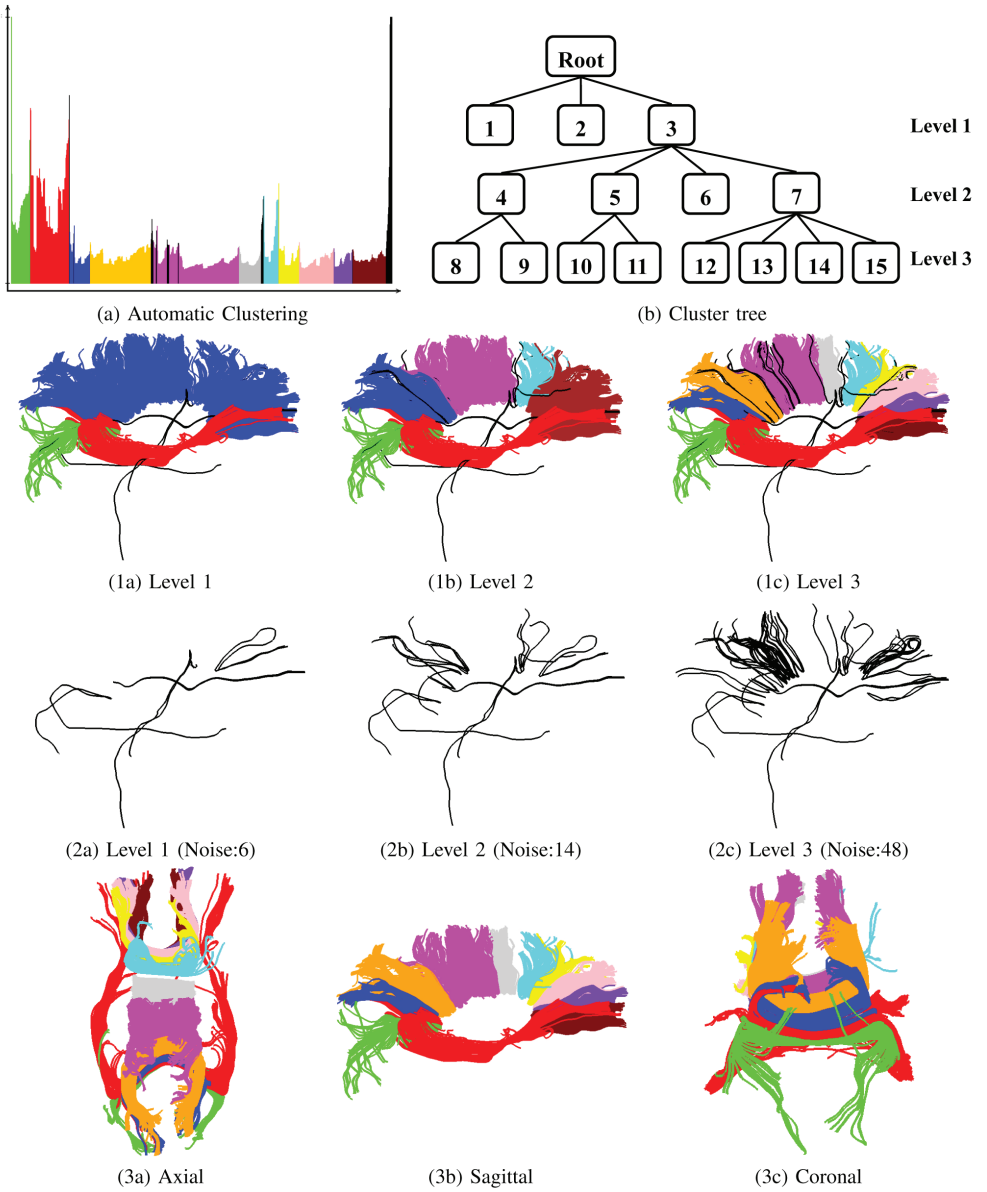
Also, at every level, noise is detected in data set 2, which comprising 1100 fibers. The first level clustering detects 6 noise fibers; this adds up to 14 at the second level. Finally we detect a total of 48 noise fibers or outliers. These numbers reflect, on the one hand, the degree of irregularity in the fiber data set, indicating erroneous fiber tracking due to wrong seed points, experimental noise, or partial volume effects. On the other hand, as mentioned, these numbers may also be influenced by the special definition of noise applied.

Our approach of automatic fiber clustering needs 4 parameters to be specified. For the algorithm OPTICS *MinPts* and *eps* must be given. For Tree Clustering, *Ratio* and *MinSize* are needed.

All calculations in this paper with OPTICS are performed with the same parametrization: *MinPts*=10 and *eps*=30. This insensitivity of OPTICS to its parameters is discussed by Ankerst et al. (1999), and greatly simplifies the application of OPTICS. Ankers et al. show that there is a broad range of possible parameters for which we can see the same clustering structure of the data in the reachability plots. Important is that *eps* is sufficiently large. The smaller we define *eps*, the more objects have an UNDEFINED reachability distance. Consequently, clusters with low density may disappear.

For Tree Clustering, the two parameters are more versatile and their selection should be guided by the user's intention: *MinSize* constrains the resolution or scale of clustering, preventing very small clusters. *Ratio* determines whether a flat or more hierarchical clustering should be detected. The larger *Ratio*, the more

Figure 13. The automatic fiber clustering result for Corpus Callosum. (a):Automatic clustering of fibers with parameters ratio=0.7, MinSize=30. (b): The Cluster tree of Corpus Callosum. (1a-1c): Clustering results at three levels respectively. (2a-2c): The noise fibers at different levels. (3a-3c): The different views of clustering results at level 3 (without noise fibers).



valleys in the reachability plot is accepted as clusters, enabling the detection of the hierarchy within the reachability plot.

CONCLUSION

We present a first application of hierarchical density-based clustering to the fiber clustering problem. A novel similarity measure is introduced, and its efficiency is increased by a lower bounding technique. The algorithm OPTICS is applied to calculate reachability plots. These plots present the cluster structure of the fiber data in a transparent manner. They are the basis for final cluster detection. We introduce for this purpose a novel interactive and an automatic cluster extraction method. Our methods compare well with existing techniques and show advantages with respect to transparency, flexibility, effectivity, efficiency and outlier robustness. The approach is tested successfully on several synthetic and real data sets.

REFERENCES

Aach, J., & Church, G. (2001). Aligning gene expression time series with time warping algorithms. *Bioinformatics (Oxford, England)*, *17*, 495–508. doi:10.1093/bioinformatics/17.6.495

Ankerst, M., Breunig, M. M., Kriegel, H. P., & Sander, J. (1999). OPTICS: Ordering Points to Identify the Clustering Structure. In *Proceedings of the ACM SIGMOD*, Philadelphia (pp. 49-60).

Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. (2000). LOF: Identifying Density-Based Local Outliers. In *Proceedings of the ACM SIGMOD 2000 Int. Conf. Management of Data*, Dallas, TX (Vol. 29, No. 2, pp. 93-104).

Brun, A., Park, H. J., Knutsson, H., & Westin, C. F. (2003). Coloring of DTMRI fiber traces using Laplacian eigenmaps. In *Proceedings of the Ninth Int. Conf. on Computer Aided Systems Theory* (Vol. 2809, pp. 564-572). Berlin: Springer Verlag.

Catani, M., Howard, R. J., Pajevic, S., & Jones, D. K. (2002). Virtual in vivo interactive dissection of white matter fasciculi in the human brain. *NeuroImage*, *17*, 77–94. doi:10.1006/nimg.2002.1136

Corouge, I., Gouttard, S., & Gerig, G. (2004). Towards a shape model of white matter fiber bundles using diffusion tensor MRI. In *Proceedings of the IEEE Int. Symposium on Biomedical Imaging*, Arlington, VA (pp. 344-347).

Damoiseaux, J. S., Smith, S. M., Witter, M. P., Sanz-Arigita, E. J., Barkhof, F., & Scheltens, P. (2009). White Matter Tract Integrity in Aging and Alzheimer's Disease. *Human Brain Mapping*, *30*, 1051–1059. doi:10.1002/hbm.20563

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series A (General)*, *39*(1), 1–31.

Ding, Z., Gore, J., & Anderson, A. (2003). Classification and quantification of neuronal fiber pathways using diffusion tensor MRI. *Magnetic Resonance in Medicine*, *49*, 716–721. doi:10.1002/mrm.10415

Dom, B. E. (2001). *An information-theoretic external cluster-validity measure* (Tech. Rep. No. RJ10219). IBM.

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second Int. Conf. on Knowledge Discovery and Data Mining* (pp. 226-231). Portland, OR: AAAI Press.

Gerig, G., Gouttard, S., & Corouge, I. (2004). Analysis of brain white matter via fiber tract modelling. In *Proceedings of the 26th Ann. Int. Conf. IEEE Engineering in Medicine and Biology Society (EMBS)*, San Francisco (pp. 4421-4424).

Hagmann, P., Cammoun, L., Gigandet, X., Meuli, R., Honey, C. J., Wedeen, V. J., & Sporns, O. (2008). Mapping the structural core of the human cerebral cortex. *PLoS Biology*, *6*(7), e159. doi:10.1371/journal.pbio.0060159

Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2001). On clustering validation techniques. *Journal of Intelligent Information Systems*, *17*(2-3), 107–145. doi:10.1023/A:1012801612483

Hansen, M., & Yu, B. (2001). Model selection and the principle of minimum description Length. *JASA*, *96*(454), 746–774.

Hinneburg, A., & Keim, D. A. (1998). An Efficient Approach to Clustering in Large Multimedia Databases with Noise. In *Proceedings of the 4th Int. Conf. on Knowledge Discovery and Data Mining*, New York (pp. 58-65).

Huang, H., Zhang, J., Wakana, S., Zhang, W., Ren, T., & Richards, L. J. (2006). White and gray matter development in human fetal, newborn and pediatric brains. *NeuroImage*, *33*, 27–38. doi:10.1016/j.neuroimage.2006.06.009

Itakura, F. (1975). Minimum Prediction Residual Principle Applied to Speech Recognition. In *Proceedings of the IEEE Trans. Acoustics, Speech, and Signal Proc.* (Vol. ASSP-23, pp. 52-72).

Jagadish, B., Yi, H., & Faloutsos, C. (1998). Efficient retrieval of similar time sequences under time warping. In *Proceedings of the 14th Int. Conf. on Data Engineering*, Orlando, FL (pp. 201-220).

Jain, A. K., & Dubes, R. C. (1988). *Algorithms for Clustering Data*. Upper Saddle River, NJ: Prentice-Hall.

Jianu, R., Demiralp, C., & Laidlaw, D. H. (2009). Exploring 3D DTI fiber tracts with linked 2D representations. *IEEE Transactions on Visualization and Computer Graphics*, *15*(6), 1449–1456. doi:10.1109/TVCG.2009.141

Keogh, E. (2002). Exact Indexing of Dynamic Time Warping. In *Proceedings of the 28th Int. Conf. on Very Large Data Bases*, Hong Kong (pp. 406-417).

Keogh, E., & Pazzani, M. (2001). Derivative Dynamic Time Warping. In *Proceedings of the First SIAM Int. Conf. on Data Mining*, Chicago (pp. 5-7).

Klein, J., Bittihn, P., Ledochowitsch, P., Hahn, H. K., Konrad, O., Rexilius, J., & Peitgen, H. O. (2007). Grid-Based Spectral Fiber Clustering. In *Proceedings of the SPIE 6509* (p. 65091E). doi:10.1117/12.706242

Law, G. Y., & Grossman, R. (2005). Application of diffusion tensor MR imaging in multiple sclerosis. *Annals of the New York Academy of Sciences*, *1064*, 202–219. doi:10.1196/annals.1340.039

Maddah, M., Grimson, W., & Warfield, S. (2006). Statistical modeling and EM clustering of white matter fiber tracts. In *Proceedings of the IEEE Int. Symp. on Biomedical Imaging*, Arlington, VA (pp. 53-56).

Maddah, M., Mewes, A., Haker, S., Grimson, W. E. L., & Warfield, S. (2005). Automated atlas-based clustering of white matter fiber tracts from DTMRI. In *Proceedings of the Int. Conf. on Medical Image Computing and Computer Assisted Intervention*, Palm Springs, CA (pp. 188-195).

Moberts, B., Vilanova, A., & van Wijk, J. J. (2005). Evaluation of Fiber Clustering Methods for diffusion tensor imaging. In *Proceedings of the IEEE Visualization*, Minneapolis, MN (pp. 65-72).

Mori, S. (2007). *Introduction to Diffusion Tensor Imaging*. New York: Elsevier.

Mori, S., Wakana, S., Nagae-Poetscher, L. M., & van Zijl, P. C. M. (2005). *MRI Atlas of Human White Matter*. Amsterdam, The Netherlands: Elsevier.

Ng, A. Y., Jordan, M. I., & Weiss, Y. (2001). On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Proc* (*Vol. 14*, pp. 849–856). Vancouver, BC, Canada: Systems.

O'Donnell, L. J., Kubicki, M., Shenton, M. E., Grimson, W. E. L., & Westin, C. F. (2006). A method for clustering white matter fiber tracts. *AJNR. American Journal of Neuroradiology*, *27*(5), 1032–1036.

O'Donnell, L. J., & Westin, C. F. (2007). Automatic Tractography Segmentation Using a Highdimensional White Matter Atlas. *IEEE Transactions on Medical Imaging*, *26*(11), 1562–1575. doi:10.1109/TMI.2007.906785

Park, H. J., Westin, C. F., Kubicki, M., Maier, S. E., Niznikiewicz, M., & Baer, A. (2004). White matter hemisphere asymmetries in healthy subjects and in schizophrenia: A diffusion tensor MRI study. *NeuroImage*, *24*, 213–223. doi:10.1016/j.neuroimage.2004.04.036

Sakoe, H., & Chiba, S. (1978). Dynamic Programming Optimization for Spoken Word Recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *26*, 623–625. doi:10.1109/TASSP.1978.1163055

Salvador, S., & Chan, P. (2004). FastDTW: Toward accurate dynamic time warping in linear time and space. In *Proceedings of the 3rd Workshop On Mining Temporal and Sequential Data (ACM KDD)*, Seattle, WA (pp. 22-25).

Sander, J., Qin, X., Lu, Z., Niu, N., & Kovarsky, A. (2003). Automatic Extraction of Clusters from Hierarchical Clustering Representations. In *Proceedings of the 7th Pacific-Asia Conf. on Knowledge Discovery and Data Mining*, South Korea (pp. 75-87).

Sankoff, D., & Kruskal, J. B. (1983). *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*. Reading, MA: Addison-Wesley.

Shao, J., Hahn, K., Yang, Q., Boehm, C., Wohlschlaeger, A., Myers, N., & Plant, C. (2010). Combining Time Series Similarity with Density-based Clustering to Identify Fiber Bundles in the Human Brain. In *Proceedings of the Int. Conf. on Data Mining (ICDM), Workshop on Biological Data Mining and its Applications in Healthcare*.

Wang, X., Wirth, A., & Wang, L. (2007). Structure-based statistical features and multivariate time series clustering. In *Proceedings of the 7th IEEE ICDM*, Omaha, NE (pp. 351-360).

Xia, Y., Turken, U., Whitfield-Gabrieli, S. L., & Gabrieli, J. D. (2005). Knowledge-based classification of neuronal fibers in entire brain. In *Proceedings of the Int. Conf. on Medical Image Computing and Computer Assisted Intervention* (pp. 205-212). New York: Springer.

Xu, X., Ester, M., Kriegel, H. P., & Sander, J. (1998). A Distribution-Based Clustering Algorithm for Mining in Large Spatial Databases. In *Proceedings of the 14th Int. Conf. on Data Engineering*, Los Alamitos, CA (pp. 324-331). Washington, DC: IEEE Computer Society Press.

Zhang, S., Demiralp, C., & Laidlaw, D. H. (2003). Visualizing diffusion tensor MR images using streamtubes and streamsurfaces. *IEEE Transactions on Visualization and Computer Graphics*, *9*(4), 454–462. doi:10.1109/TVCG.2003.1260740

Zhang, S., & Laidlaw, D. H. (2005). DTI fiber clustering and cross-subject cluster analysis. In: International Society for Magnetic Resonance in Medicine (Ed.), *Proceedings of the 13th Scientific Meeting ISMRM*, Miami, FL (pp. 2727).

Junming Shao is currently a PhD student of Insititute of informatics at University of Munich, Germany (LMU). He received his bachelor and master degree in Computer Science from Northwest A&F University, China, in 2005 and 2008 respectively. Currently, he is also working with neuroimaging group in the department of Neuroradiology, Klinikum rechts der Isar of the Technical University to investigate the relationship between structural and functional connectivity in Alzheimer's disease using diffusion tensor imaging and fMRI. Besides his basic research of data mining, he has been applying data mining techniques in diverse fields, such as biology, neuroscience and environment.

Klaus Hahn has a permanent scientist-position at the institute for Biomathematics and Biometrics of the Helmholtz-Zentrum Munich, Germany. His main scientific contributions are in the fields : Nuclear Physics, Radiation Biology and Neuro Imaging. After studying theoretical physics at the LMU University Munich he achieved his Doctor degree at University Tübingen. His contributions cover theoretical, mathematical, numerical and statistical investigations and led to about 70 refereed papers. He is also engaged in the education of young scientists and CoEditor of the International Journal of Biomathematics and Biostatistics.

Qinli Yang obtained her bachelor degree from Northwest A&F University in the subject Environmental Science in 2007. Currently, she is pursuing her PhD in the field of Environment Engineering in the University of Edinburgh. Her research interests include water engineering, flood risk control and data mining.

Afra Maria Wohlschläger currently is the head of the neuroimaging group in the department of Neuroradiology, Klinikum rechts der Isar of the Technical University in Munich, Germany. She received her master and PhD degree in Physics at University of Cologne, Germany. Her research fields include resting state activity of the brain, functional connectivity analysis, multimodal integration: functional magnetic resonance imaging (fMRI), Diffusion Tensor Imaging (DTI), Voxel Based Morphometry (VBM), etc.

Christian Boehm is currently professor of informatics at Ludwig-Maximilians-Universit Universität München, Germany (LMU) and visiting professor at Florida State University, Tallahassee (FSU). His previous affiliations include Technische Universität München (TUM) and the University for Medical Informatics and Technology in Innsbruck, Austria (UMIT). He has currently more than 90 reviewed publications in database research (with focus on indexing) and data mining (with focus on clustering). He holds one patent and has received the SIGMOD best paper award 1997 and the best paper honorable mention award at the SIAM Int. Conf. on Data Mining (SDM) 2008.

Nicholas Myers studied Neuroscience at Columbia University, where he assisted in using brain imaging techniques to study human language comprehension. While working at the Technical University in Munich he has collaborated on using brain imaging to identify the relationship between structural and functional alterations and attentional impairments in early Alzheimer's disease using diffusion tensor imaging, resting state fMRI, and PET. He is currently a student of Neuroscience at the University of Oxford, working on brain imaging studies of attention and memory.

Claudia Plant currently is visiting professor at the department of Scientific Computing at the Florida State University sponsored by a Feodor Lynen fellowship of the Alexander von Humboldt Foundation. She received her PhD in biomedical informatics in 2004 from the University of Health Sciences Medical Informatics and Technology (UMIT), Austria. She has been working on several biomedical data mining projects at UMIT, University of Munich and Technische Universität München, Germany with applications in metabolomics, proteomics and neuroscience. Besides mining biomedical data she has been contributing to basic research especially in clustering, high-performance data mining and information-theoretic data mining.

APPENDIX

Remark: LB is a lower bound to DTW for spatial fibers.

For two fibers $P = (p_1, \ldots, p_m)$ and $Q = (q_1, \ldots, q_n)$, for each dimension, we have $lb(p_d, q_d) < dtw(p^d, q^d)$, $d = \{1, 2, 3\}$ (Yi, Jagadish, & Faloutsos, 1998). We want to prove

$DTW(P, Q) > LB(P, Q)$

This is shown in the following.

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