A Novel Similarity Measure for Fiber Clustering using Longest Common Subsequence

Christian Böhm
University of Munich
boehm@dbs.ifi.lmu.de

Jing Feng
University of Munich
feng@dbs.ifi.lmu.de

Xiao He
University of Munich
he@dbs.ifi.lmu.de

Son T. Mai
University of Munich
mtson@dbs.ifi.lmu.de

Claudia Plant
Florida State University
cplant@fsu.edu

Junming Shao
University of Munich
shao@dbs.ifi.lmu.de

ABSTRACT

Diffusion tensor imaging (DTI) is an MRI-based technology in neuroscience which provides a non-invasive way to explore the white matter fiber tracks in the human brain. From DTI, thousands of fibers can be extracted, and thus need to be clustered automatically into anatomically meaningful bundles for further use. In this paper, we focus on the essential question how to provide an efficient and effective similarity measure for the fiber clustering problem. Our novel similarity measure is based on the adapted Longest Common Subsequence method to measure shape similarity between fibers. Moreover, the distance between start and end points of a pair of fibers is also included with the shape similarity to form a unified and flexible fiber similarity measure which can effectively capture the similarity between fibers in the same bundles even in noisy conditions. To enhance the efficiency, the lower bounding technique is used to restrict the comparison of two fibers thus saving computational cost. Our new similarity measure is used together with density-based clustering algorithm to segment fibers into groups. Experiments on synthetic and real data sets show the efficiency and effectiveness of our approach compared to other distance-based techniques, namely Dynamic Time Warping (DTW), Mean of Closest Point (MCP) and Hausdorf (HDD) distance.

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Data mining

General Terms

Algorithms

Keywords

Fiber Clustering, Longest Common Subsequence, Diffusion Tensor Imaging

DMMH'11, August 21, 2011, San Diego, California, USA.

1. INTRODUCTION

Diffusion tensor imaging (DTI) is a structural magnetic resonance imaging (MRI) which helps to measure microscopic movement of water in the brain, and it has been used to explore organization and integrity of white matter structures of human brain in vivo. From DTI data, the white matter tracts can be reconstructed via a process called tractography which estimates white matter tract trajectories by following likely track directions [13]. This information can be used in surgical planning and in study anatomical connectivity, brain changes and mental disorders [12].

After tractography, we obtain thousands of fiber trajectories, and they need to be grouped into anatomical meaningful structures before being used. One common method is based on the knowledge of experts and is referred to as *virtual dissection* [4]. This method interactively selects fibers passing through some manually defined region of interests (ROIs). This process is highly flexible and can help to detect anatomically meaningful bundles with very different shapes. However, it is very time consuming and may be biased by subjective points of view of the experts. Therefore, automatic clustering of white matter fiber tracks is an interesting alternative for many applications [5, 6, 3].

One essential problem in automatic fiber clustering is to provide a similarity measure for a pair of fibers. Two fibers are usually considered as similar if they are separated by a small distance, have comparable length and similar shape [6]. However, these criteria may be not sufficient. Two fibers with different shapes, for example, can be grouped into the same bundle if they start and end at the same regions [3]. Moreover, due to the scanning process of DTI, each fiber may contain some amount of noise, which can affect the similarity between them [4]. Although, there are many proposed techniques in literature [5, 14, 6, 15, 11], much efforts are going on to find out more effective and efficient procedures.

Among various techniques, distance based similarity measure ones like Hausdorf distance (HDD), Mean of Closest Point distance (MCP) [5] and Dynamic Time Warping (DTW) [14] are widely used. However, their point-to-point distance measure mechanism is sensitive to noise, which affects the final distance similarity between pairs of fibers. Moreover, only the final distance between two fibers is not enough to tell whether they have a similar shape or they are separated by a small distance. Thus, this limits their ability to effectively group fibers into meaningful bundles. Besides, with

HDD, MCP and DTW, their quadratic time complexity is undesirable, especially for large data sets.

In this paper, we propose a novel similarity measure for fiber clustering which is based on a modified Longest Common Subsequence (LCSS) technique. LCSS is specially adapted to points in three-dimensional space to measure shape similarity, and is less sensitive to noise than other distance based methods. In addition, the distance between start and end points of a pair of fibers, which is referred to as distance similarity, is also incorporated to effectively capture the complex notion of fiber similarity. As a result, our approach provides an effective and flexible way to the similarity between fibers. Besides, to enhance efficiency of our algorithm, we use a lower bounding technique to terminate the comparison of fibers as soon as it can be omitted, thus saving computational cost. Our similarity measure is used together with a density-based clustering algorithm DBSCAN [9] to group fibers into tracks and eliminate outliers. Experiments on synthetic and real data sets show that our similarity measure is superior to rival techniques namely DTW, MCP and HDD in terms of both efficiency and effectiveness.

2. BACKGROUND AND RELATED WORK

Providing a similarity measure for a pair of fibers is an essential problem for automatic fiber clustering. In the early work by Brunt et al. [3], two fibers are considered as similar if their start and end points are close together. However, this assumption is not always reasonable, since not all fibers in a same bundle start and end in the same regions [4]. Furthermore, it also ignores the shape similarity between fibers. Ding et al. [6] defined the similarity by using the mean Euclidean distance and segment ratio between their corresponding segments. However, how to find a good corresponding segment remains a question. This algorithm also ignores the important roles of the start and end points of fibers [4]. Zhang et al. [18] used average distance from points in the shorter fiber to their closest points in the longer one (if they are larger than a predefined threshold). Later, Courouge et al. [5] introduced three similarity functions: closest point distance, mean of closest point distance (MCP) and Hausdorff distance (HDD), which are used widely. These functions measure similarity by using the point pairs distance between two fibers. Recently, Shao et al. [14] extended this idea by using Dynamic Time Warping (DTW) due to its flexibility with varying length fibers. However, these point-to-point distance-based measure techniques, similar to Zhang et al. [18], are sensitive to noise. Moreover, their distance measure mechanism is not strong enough to tell whether two fibers have similar shape or not, thus limiting their effectiveness. Besides, MCP, HDD and DTW have quadratic time complexity, which is undesirable, especially for large fiber data sets.

Assume that we have two fibers $P = (p_1, \cdots, p_n)$ and $Q = (q_1, \cdots, q_n)$. MCP and HDD are defined as follow [5]:

$$d_{MCP}(P,Q) = avg(d_{mcp}(P,Q), d_{mcp}(Q,P))$$
$$d_{HDD}(P,Q) = max(d_{hdd}(P,Q), d_{hdd}(Q,P))$$

where

$$d_{mcp}(P,Q) = avg_{p_i \in P}\{min_{q_j \in Q}\|p_i - q_j\|\}$$
$$d_{hdd}(P,Q) = max_{p_i \in P}\{min_{q_j \in Q}\|p_i - q_j\|\}$$

Let P_i be the first i points of P. DTW distance between P and Q can be defined as follows [14]:

$$d_{DTW}(P,Q) = \frac{d_{dtw}}{K}$$

where K is the length of warping path [10] and d_{dtw} is defined recursively as follows [17]:

$$d_{dtw}(P,Q) = \|p_n - q_m\| + min(d_{dtw}(P_{n-1}, Q_m), d_{dtw}(P_n, Q_{m-1}), d_{dtw}(P_{n-1}, Q_{m-1}))$$

Another problem in fiber clustering is choosing a suitable clustering algorithm. There are many clustering algorithms such as traditional EM clustering [8], k-nearest neighbors method [5], spectral clustering [3, 13], hierarchical clustering [18] and density-based clustering [14]. Of all mentioned above, we are especially interested in density-based clustering algorithms due to their abilities to discover clusters with arbitrary shape and to deal with outliers. Therefore, in this paper, we decided to use DBSCAN [9], a well-known density-based clustering algorithm, together with our new similarity measure to group fibers into bundles.

3. FIBER SIMILARITY MEASURE

After deterministic tractography, a fiber is represented as an ordered set of points with different arc lengths and different number of points in 3D space. To quantify fiber similarity, we use the distance between start and end points of a pair of fibers as the distance similarity between them. Distance similarity is then combined with the use of modified Longest Common Subsequence (LCSS) technique as a shape similarity measure in a weighted scheme to provide a unified, efficient and flexible way to capture the similarity between fibers.

In the following sections, firstly we introduce LCSS technique and how to use it to compare shape similarity between two fibers. Secondly, we present lower bounding technique to reduce the computational cost. Lastly, we discuss the final similarity measure function and its characteristics.

3.1 Longest Common Subsequence

Longest common subsequence [2] is a classic, famous and well-studied computer science problem, which finds the longest subsequence common to all sequences in a set of sequences and is applied in many fields such as bioinformatics [1] and time series data mining [17]. However, to the best of our knowledge, it has never been applied to the fiber clustering problem so far.

Assume that we have two sequences, $X = (x_1, \ldots, x_n)$ and $Y = (y_1, \ldots, y_m)$. We need to find longest common subsequence $Z = (z_1, \ldots, z_k)$ of X and Y, where there exists strictly increasing sequence $i = (i_1, \ldots, i_k)$ and $j = (j_1, \ldots, j_k)$ of indices of X and Y such that for all $l = 1, \ldots, k$, $x_{i_l} = z_l$ and $y_{j_l} = z_l$. For example, if $X =$ (1,2,1,3,2,2,1,3,1) and $Y =$ (1,1,2,2,2,3,3), then $Z = (1, 1, 2, 2, 3)$.

3.2 Shape similarity of two fibers

Assume that we have two fibers $A = (a_1, \ldots, a_n)$ and $B = (b_1, \ldots, b_m)$ in 3D. To measure the shape similarity between A and B, we build an envelope around A and then compare this envelope with B. If B is in the envelope of A then they have similar shape. This kind of comparison provides a new view for fiber shape similarity, which differs from the distance-based methods like DTW, MCP and

HDD. Consider Figure 1, by distance-based mechanism, we cannot know whether the shape of fiber B or C is similar to A or not, because $Dist(A,B) \approx Dist(A,C)$. However, the envelope scheme successfully discovers that the shape of C is more similar to A than B, because a large part of C lies inside the envelope of A.

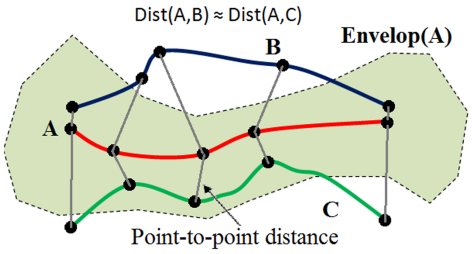

Figure 1: Both fiber B and C are similar to fiber A by point-to-point distance-based similarity measure.

This envelope scheme can be simulated by considering fibers as time ordered sequences of points and using an extended LCSS model to measure their shape similarity.

Definition 1. Two fibers A and B are close to each other at position i and j respectively if the point coordinates at i and j are not different more than a similarity threshold ϵ.

$$a_i \text{ is close to } b_j \Leftrightarrow |a_i(x) - b_j(x)| \le \epsilon \\ \text{and } |a_i(y) - b_j(y)| \le \epsilon \text{ and } |a_i(z) - b_j(z)| \le \epsilon$$

Let $A_i = (a_1, \ldots, a_i)$ be the subsequence from position 1 to i (or prefix with length i) of A.

Definition 2. Given a time constraint δ and a similarity threshold ϵ, the length of longest common subsequence of two fibers A and B, $LCSS_{\delta,\epsilon}(A_n, B_m)$ (or $LCSS_{\delta,\epsilon}(A,B)$), can be defined as follows:

$$LCSS_{\delta,\epsilon}(A_n, B_m) = \begin{cases} 0 & \text{if } A \text{ or } B \text{ is empty} \\ 1 + LCSS_{\delta,\epsilon}(A_{n-1}, B_{m-1}) & \\ \quad \text{if } a_n \text{ is close to } b_m & \\ \quad \text{and } |n-m| \le \delta & \\ max(LCSS_{\delta,\epsilon}(A_{n-1}, B), & \\ \quad LCSS_{\delta,\epsilon}(A, B_{m-1})) & \text{otherwise} \end{cases}$$

where δ constrains the matching regions in time to avoid two sequences to be compared at too far away positions, which may be nonsense and unnecessary. LCSS can be calculated by using dynamic programming approach [2, 17] to construct a cost matrix $M_{n \times m}$, where the value of $M_{i,j}$ can be calculated by the values of its adjacent cells. The time complexity of this LCSS model is thus $O(\delta(n+m))$ [17].

Figure 2 illustrates the calculation of LCSS for two fibers A and B in 1D. The time constrain δ limits the comparison of the point a_i with the points b_j with j in $[i-\delta, i+\delta]$ only. Thus, it improves the efficiency of the algorithm. The point b_j is matched with a_i if and only if it lies inside the ϵ-circle of a_i.

The value of LCSS is unbounded and depends on the length of sequences. Therefore, in order to compare sequences of variable length, we need to normalize the cost [17].

Definition 3. Given a time constrain δ and a similarity threshold ϵ, the shape similarity of A and B can be calculated from $LCSS_{\delta,\epsilon}(A,B)$ as follows:

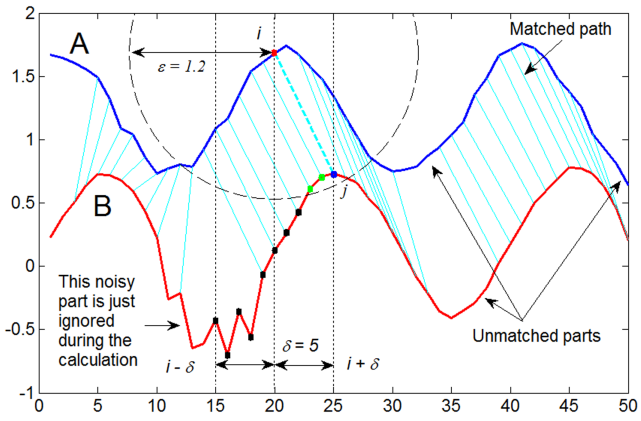

Figure 2: The comparison of two fiber A and B by LCSS technique in 1D with the time constraint δ and similarity threshold ϵ. For each point a_i in fiber A, only one point b_j in fiber B (with j in $[i-\delta, i+\delta]$), which lies inside ϵ-circle of a_i, can be matched with a_i.

$$Shape_{\delta,\epsilon}(A,B) = 1 - \frac{LCSS_{\delta,\epsilon}(A,B)}{min(n,m)}$$

We call $Shape_{\delta,\epsilon}(A,B)$ the shape similarity of two fibers A and B with respect to the time constraint δ and the similarity threshold ϵ. The smaller the value of $Shape_{\delta,\epsilon}(A,B)$ is, the more similar the shape of two fibers A and B are.

3.3 Lower bounding distance

We use the lower bounding distance [17] to further enhance the efficiency by eliminating the unnecessary calculations of $LCSS_{\delta,\epsilon}$.

The lower bounding distance of $LCSS_{\delta,\epsilon}(A,B)$ of two equal length fibers can be calculated by using the Minimum Bounding Envelope of A ($MBE_{\delta,\epsilon}(A)$) with respect to the time constraint δ and the similarity threshold ϵ. To make it easier to reader, we assume that A and B are now 1D fibers. However, the notion of the $MBE_{\delta,\epsilon}$ can be trivially extended to 3D [17].

$$Envlow \le MBE_{\delta,\epsilon}(A) \le Envhigh$$

where

$$\begin{cases} Envhigh_i = max(a_j) + \epsilon & \forall j, |i-j| \le \delta \\ Envlow_i = min(a_j) + \epsilon & \forall j, |i-j| \le \delta \end{cases}$$

The length of longest common subsequence between B and $MBE_{\delta,\epsilon}(A)$ is defined as follows:

$$LCSS(MBE_{\delta,\epsilon}(A), B) = \sum_{i=1}^{n} \begin{cases} 1 \text{ if } a_i \text{ in envelope} \\ 0 \text{ otherwise} \end{cases}$$

In Figure 3, only the parts of B that lie inside $MBE_{\delta,\epsilon}$ of A are used to calculate the upper bound of $LCSS_{\delta,\epsilon}(A,B)$ or lower bound of $Shape_{\delta,\epsilon}(A,B)$.

Lemma 1. $LCSS(MBE_{\delta,\epsilon}(A), B)$ is the upper bound of $LCSS_{\delta,\epsilon}(A,B)$ and thus $Shape_{\delta,\epsilon}(MBE_{\delta,\epsilon}(A), B)$ is the lower bound of $Shape_{\delta,\epsilon}(A,B)$ for any time series A and B [17].

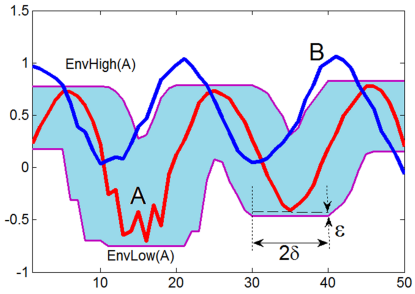

Figure 3: The Minimum Bounding Envelope (MBE) of fiber A with respect to the time constraint δ and similarity threshold ϵ is used to calculate lower bounding distance with fiber B. Only the parts of B that lie inside $MBE_{\delta,\epsilon}$ of A can be matched during the computation of LCSS.

To deal with the varying lengths of fibers, we extend the concept of this lower bounding by building an envelope for the longer fiber and compare this envelope with the shorter fiber. It is simply and straight forward to see that the lower bounding property still hold. That means: $LB_{\delta,\epsilon}(A,B) = Shape_{\delta,\epsilon}(MBE_{\delta,\epsilon}(A),B) \leq Shape_{\delta,\epsilon}(A,B)$, assuming that A is longer than B.

This lower bounding distance is specially important when we do the Range Query Search within a density-based clustering algorithm to find fibers in set D which are similar to a given query fiber p and similarity threshold ϵ (Figure 4).

```
Function rangeQuery(p, ε, D)
    Neighbors = {}
    for all object q in D
        LB_dist = LB(p, q)
        if (LB_dist < ε)
            if (Dist(p,q) < ε)
                Add q to Neighbors
            endif
        endif
    endfor
Return Neighbors
```

Figure 4: Pseudocode for ϵ-range query using lower bounding distance.

3.4 Distance similarity of two fibers

Given two fibers A and B with the start points p_A and p_B and the end points q_A and q_B. We define the distance similarity between A and B as follows:

$$Dist(A,B) = \|p_A - p_B\| + \|q_A - q_B\|$$

The distance similarity measures how close the two fibers start and end. Though, the Gaussian function $e^{-\frac{x^2}{2\sigma^2}}$ is preferred [3] to measure similarity, the difference between the Gaussian and the linear function is negligible in our experiments. Thus, we chose the linear function. However, despite of the used functions, data should be normalized to acquire the homogeneousness among different data sets. This will make the selection of parameters much easier.

3.5 Unified fiber similarity measure

We define the similarity between two fibers A and B as a weighted sum of their shape and distance similarities.

$$Sim_{\delta,\epsilon,\alpha}(A,B) = \alpha \cdot Shape_{\delta,\epsilon}(A,B) + (1-\alpha) \cdot Dist(A,B)$$

where α lying between $[0,1]$ is used to control the balance between the shape and distance similarities.

The shape similarity includes the notion of fiber similarity of Ding et al. [6]. And the distance similarity captures the notion of fiber similarity of Brun et al. [3], which is also close to the use of manual ROIs of Catani et al. [4]. $Sim_{\delta,\epsilon,\alpha}(A,B)$ unifies the notions of fiber similarity proposed so far. We will show that, it is an effective method to measure similarity between fibers.

The use of α provides not only a unified notion but also a flexible way to enhance the effectiveness of fiber similarity. When the two fibers A and B belong to two close bundles, for example, Arcuate and Superior Longitudinal Fasciculus in Figure 5, it may be hard to distinguish them under $Shape_{\delta,\epsilon}$. In this case, the use of α can improve the effectiveness of the algorithm by making the distance similarity more important.

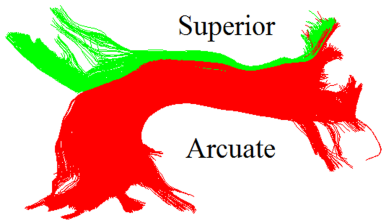

Figure 5: Two bundles Arcuate and Superior Longitudinal Fasciculus are close to each other and hard to distinguish.

3.6 Other important characteristics of LCSS

Due to the process of DTI tractography, the fibers may contain noise, which can affect the similarity between them [4]. Consider Figure 6, assume that A and B in the upper part are two real fibers and in the lower parts are noisy fibers. We calculate DTW, HDD, MCP and $SIM_{10,0.2,1}$ to see the effect of noises in each measure. We can see that, SIM is more robust to noise than other methods. The value of SIM didn't change, because the noisy part is ignored in the calculation of SIM.

Another problem occurs when we try to consider a spatial fiber as an time ordered sequence of points. Depending on the way we write down the value of A (normal or reverse order), the similarities between A and B shall be very different. This phenomenon often happens in DTI tractography, when one fiber is recognized in a direction contrary to the rest in a group. To overcome this problem, we use 2-phases approach: first calculate the similarity between $(A,\ B)$ and $(Rev(A),B)$ and then choose the smallest result, as follows:

$$Sim2_{\delta,\epsilon,\alpha}(A,B) = min(Sim_{\delta,\epsilon,\alpha}(A,B), Sim_{\delta,\epsilon,\alpha}(Rev(A),B))$$

Figure 7 a) shows the result of the clustering using 1-phase approach, there are many false direction fibers which couldn't be correctly grouped. Figure 7 b) shows perfect cluster result (exactly the same as gold standard) if we use 2-phases approaches.

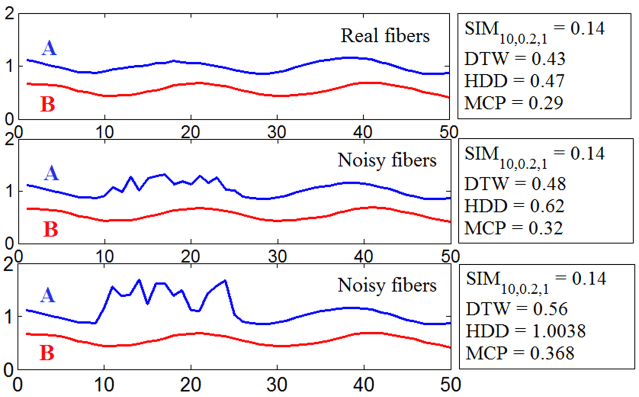

Figure 6: The effect of noise on different similarity measure functions between two time series A and B. SIM is more robust to noise than other methods.

Figure 7: The results of clustering using 1-phase similarity measure (10 clusters) a) and 2-phases similarity measure (5 clusters) b). The 2-phases approach produces the gold standard exactly.

MCP and HDD are not affected by this phenomenon because they measure the similarity by using the distance between each point in fiber A and their closest point in fiber B. DTW and LCSS technique, on the contrary, are affected by this phenomenon, since they consider fibers as time ordered sequences and constrain the comparison of them in an incremental way.

4. DENSITY-BASED CLUSTERING

In density-based clustering methods, clusters are considered as areas of high object density and separated by areas of low object density. This notion has several attractive benefits. Users do not need to specify the number of clusters. It can detect clusters of arbitrary shape, and is robust to noise and outliers. Among many proposed approaches namely DENCLUE, OPTICS and DBCLASD, DBSCAN [9] is the well-known one. It is based on the idea that each point of a cluster has to contain at least *minpts* points within its *eps*-neighborhood. The algorithm DBSCAN maintains the seed list which contains a set of seed objects S for cluster expansion. Figure 8 shows the pseudo code for DBSCAN clustering algorithm.

DBSCAN requires 2 parameters: *minpts* and *eps*, which can be estimated from a sorted k-dist graph [9].

5. EXPERIMENTS

In this section, we present our experiments on synthetic and real data sets to prove the efficiency and effectiveness of our algorithm. Part of the real data sets was extracted from

```
Function DBSCAN(D, minpts, eps)
  currentID = Fist_ID
  set all object p in D to unprocessed
  for all object p in D do
    if p is processed then continue endif
    S = rangeQuery(p, ε, D)
    if S.size < minpts then
      assign cluster id of p as noise
    else
      assign cluster id of p as currentID
      assign cluster id for all objects in S as CurrentID
      while S not empty do
        q = S.first( )
        T = rangeQuery(q, ε, D)
        if T.size >= minpts then
          for all objects r in T do
            if r is unprocessed or noise
              if r is unprocessed then
                insert r into S
              endif
              set cluster id of r as currentID
            endif
          endfor
        endif
        remove q from S
      endwhile
      currentID = Next_ID
    endif
  endfor
EndFunction
```

Figure 8: Pseudo code for DBSCAN algorithm used in our work.

the labeled data set acquired from PBC Brain Connectivity Challenge - IEEE ICDM - Fall 2009 (http://pbc.lrdc.pitt.edu/?q=2009b-home). This data set contains 250000 fibers. However, only 29029 fibers belonging to 8 famous bundles are labeled. The other data sets are provided by our experts. All of them are normalized by scaling their bounding boxes to ensure that they do not exceed the range [-1,1] on each coordinate. Our algorithm is implemented in Java together with Mathlab as visualization tools. All experiments are conducted on Laptop with CPU Core i7 1.6 Ghz, 6GB Ram.

5.1 Parameters and cluster validation

To compare the clustering results with the gold standards, we used two different measure methods. The first one is the information-theoretic external cluster-validity [7], which measures how useful the calculated cluster labels are as predictors of the gold standard cluster labels. The smaller and closer the value of DOM to the coding cost of gold standard is, the better the result is. The second one is Normalized mutual information (NMI) [16], which measures the mutual dependence of the cluster result and the gold standard. The result is in [0,1], with 0 means that the cluster result is independent with the gold standard and 1 means that the cluster result is the same as the gold standard. These methods, in contrast to others namely Rand Index or Cluster Purity, can compare results with different numbers of clusters. To make it easier to compare DOM and NMI, we also use normalized DOM cost (nDOM) by dividing the gold coding cost and the cluster coding code to obtain value in (0,1].

To ensure fairness while comparing different approaches, we use exhaustive searches over parameter spaces to find the best parameter combinations for each approach. For example, with the parameter *eps* of DBSCAN, we explore the search space from the minimum value 0.01 to 1.0 with search step 0.01 to ensure that we do not miss any good results. To avoid confusing readers with too many parameters, we

report the parameter *eps* of DBSCAN only. For the other parameters, the default values are $\delta = 50$, $\epsilon = 0.05$ and $\alpha = 0.8$ for SIM and $minpts = 5$ for DBSCAN unless otherwise stated. The detailed discussion about the parameters can be found in the last part of this section.

5.2 Effectiveness of Similarity Measure

In this section, we will compare the effectiveness of our new fiber similarity measure with some well-known techniques, namely DTW, HDD and MCP. Our algorithm (SIM) and DTW are run in 2-phases (while the algorithm from Shao et al. [14] only run in 1-phase).

The synthetic data set (Syn) contains 520 fibers in 10 clusters with 5 clusters of straight lines, 2 cluster of helices, 3 clusters of fiber-like objects and 8 outliers which are added to demonstrate the ability of the algorithms to detect outliers. Figure 9 shows the best found cluster results of SIM, DTW, MCP and HDD (outliers are always drawn in black). All techniques can recognize exactly the 10 clusters. However, only SIM can procedure exactly the gold standard. DTW and MCP fail to classify several fibers and HDD is even worse.

Figure 10 shows cluster results on a real data set. This data set (Rel) was randomly extracted from the PBC data set and contains 500 fibers belonging to 5 bundles namely Arcuate, Cingulum, Fornix, Inferior Occipitofrontal Fasciculus and Superior Longitudinal Fasciculus. Moreover, five fibers from other bundles are randomly extracted and added into this data set as outliers. Only SIM can produce gold standard exactly. Apart from minor errors, DTW detects the 5 clusters, while MCP and HDD result in 6 clusters.

Let consider another aspect of the effectiveness of similarity measure. As we know, the parameter *eps* of DBSCAN specifies the range of the core objects. Thus, it plays an important role to distinguish fiber bundles. Therefore, a better similarity measure should support the wider range of *eps* (assume that we fixed *minpts*). In this experiment, we let *eps* run in the range from 0.01 to 1.0 with step size 0.01 and count the numbers of *eps* values which result in cluster scores better than a predefined threshold (0.9 for NMI and 70% of gold standard for DOM) on the same data set used in Figure 10. As we can see from Figure 11, SIM supports the widest range of *eps*, so it is the most robust technique for fiber clustering. The results are also the same with all other data sets in our experiments.

To see how these techniques perform in noisy environments, we add 5% random Gaussian noise into the synthetic data set used in Figure 9. We also put two point outliers into two random fibers to make the data set (Synnoise) more difficult. While all other techniques were affected by noises, SIM is totally not affected. It still produced gold standard exactly. Moreover, all point-to-point distance measure techniques can not group the two fibers with local outliers, because the outliers significantly affect the comparison of pairs of fibers. However, with SIM, those points are just ignored during the comparison of fibers. Thus they have little effect on the final similarities. As a result, SIM is more robust to noise than any other distance based techniques.

5.3 Efficiency of Similarity Measure

To study the efficiency of the algorithms, we do an sequential ϵ-range search on a fiber data set with 10000 objects with LCSS, DTW, MCP and HDD. Two parameters

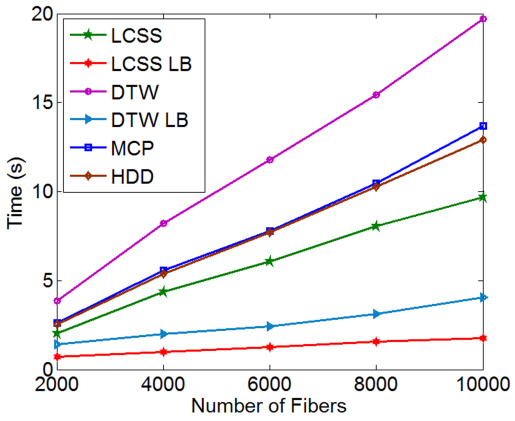

Figure 13: Comparison of efficiency between different fiber similarity measure techniques.

δ and ϵ of LCSS are set to 50 (about quarter of average fiber lengths) and 0.2 respectively. Figure 13 shows running time of range query on LCSS and DTW (with and without lower bounding distance), MCP and HDD. We can see that LCSS with lower bounding techniques outperforms DTW, MCP and HDD at all. It is easy to understand since the time constraint δ of LCSS limits the comparison paths thus saving computational cost. The tightness of lower bounding distance [10] of LCSS is also higher than of DTW [14], about 79% for LCSS and 28% for DTW in real data sets. To cluster the data set (Rel) in Figure 10, DTW finished in 49 seconds while our algorithm (SIM) in only 10 seconds. MCP and HDD completed only after 106 and 105 seconds.

Fibers	5000	15000	25000	50000
nDom	0.539	0.573	0.553	0.456
Nmi	0.956	0.935	0.937	0.931
Times(s)	101	479	1201	4853

Table 1: Running times and scores of our algorithm (SIM) on some real data sets which contain 5000 to 50000 fibers.

Table 1 shows the running times and scores of our algorithms in some real data sets extracted randomly from PBC. Only a quarter of fibers are labeled in each data set, the rest is unlabeled. We score the results only on labeled fibers. Our algorithms run very fast with good scores compared with other techniques. For example, to cluster the data set with 5000 fibers, DTW needs 1557 seconds to reach the same score as SIM, while MCP and HDD are much slower.

5.4 Fiber clustering on real data sets

Figure 14 demonstrates the clustering ability of our algorithm on 3 real data sets. These data sets are also extracted randomly from the PBC labeled data set with 8 bundles namely Arcuate, Cingulum, Fornix, Inferior Occipitofrontal Fasciculus, Superior Longitudinal Fasciculus, Forceps Major and Corticospinal. They contain 5, 6 and 8 different bundles with 500, 1200 and 1500 fibers respectively. Also 5 fibers from other groups are added into each data set as outliers. All of them are clustered exactly as gold standard. Note that, DS3 is clustered with $\alpha = 0.6$.

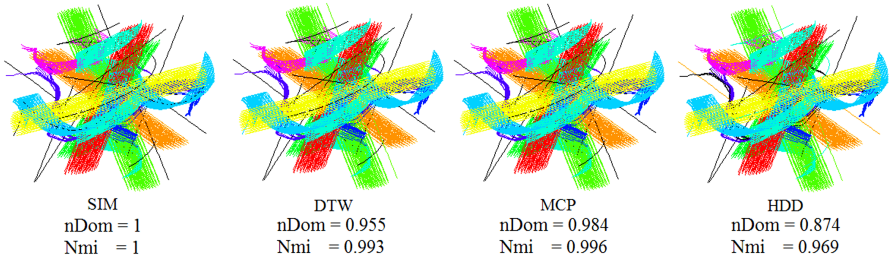

Figure 9: Effectiveness of different similarity measure techniques for synthetic data. Only SIM produced exactly the gold standard. Best results are obtained with $eps = 0.23, 0.18, 0.08, 0.25$ respectively.

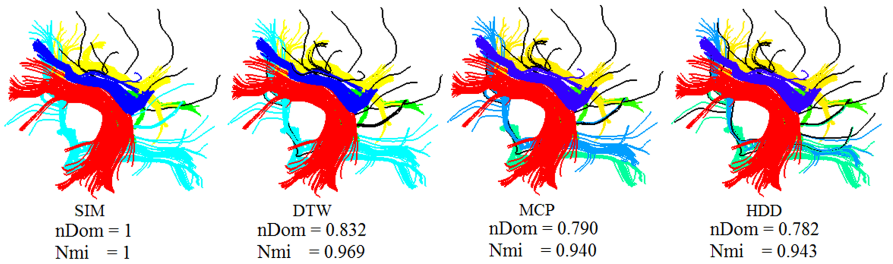

Figure 10: Effectiveness of different similarity measure techniques for real data. Only SIM produced exactly the gold standard. Best results are obtained with $eps = 0.44, 0.1, 0.06, 0.26$ respectively.

Figure 15 shows cluster results for some real data sets proposed by our experts to assess our algorithm. Although we do not have the gold standards to compare, all results were well-confirmed by our experts. DF2 was clustered with $\delta = 100$ and DF3 was clustered with $\delta = 100$, $\epsilon = 0.1$ and $\alpha = 0.5$. The data sets DF1, DF2 and DF3 are acquired from Shao et al. [14].

5.5 How to select the parameters

Although SIM is superior to other techniques in terms of efficiency and effectiveness, it requires 3 more parameters: the time constraint δ, the similarity threshold ϵ and the weight α, which may confuse us at the first glance. However, these parameters are actually easy to set up.

The first parameter δ strongly affects the efficiency rather than effectiveness of the algorithm, thus we can easily select its value. However, too small values lead to very narrow comparison ranges and decrease the flexibility of the shape similarity measure (especially for fibers with very different lengths). Conversely, too high values make the algorithm slower. In our experiments, the results depend slightly on the value of δ. So, we simply choose δ about a quarter of the average length of all fibers.

The second parameter ϵ affects slightly the effectiveness of the algorithm. However, if we choose too small value, we cannot capture shape similarity of fibers in the same group, because they shall fall outside the envelope of each other, thus they are totally different to one another under LCSS. If ϵ is too big, we could not distinguish fibers from different groups because they look similar under LCSS. Therefore, ϵ should be chosen to ensure that the envelope of one fiber can contain sufficient number of fibers (about *minpts* fibers) from the same group. The value of ϵ can be estimated if we visualize the data. In our experiments, we simply choose $\epsilon = 0.05$ for most cases.

The last parameter α plays more important role. It controls the balance between the shape and distance similarity. Thus, it affects the effectiveness of our algorithm. To handle it correctly, we need to understand clearly the correlation between α and the cluster results on each kind of data sets. Figure 16 shows the relationship between different values of α and NMI scores, DOM scores as well as their *eps* ranges (with the threshold of 0.95 for NMI and 80% for DOM) on three real data sets DS1, DS2 and DS3 in Figure 14. As we see, each data set depends on α in slightly different ways. But all of them acquire good and stable performances when $\alpha \geq 0.3$. Besides, the *eps* ranges in all data sets increase with α, which means that the larger the value of α is, the better and more flexible our algorithm is. Due to its widely acceptable range, the choice of α is also not so hard. In all our experiments, we simply set $\alpha = 0.8$ and that is enough for most of the data sets. The choice of $\alpha = 0.5$ is also acceptable.

	Syn	Synnoise	Rel	DS1	DS2	DS3
SIM	1	1	1	1	1	1
Shape	1	1	0.961	1	1	0.986
DTW	0.993	0.986	0.969	1	1	0.976
MCP	0.996	0.989	0.941	1	1	0.976
HDD	0.969	0.964	0.944	0.925	0.952	0.945

Table 2: NMI scores of some measure techniques on some data sets.

We emphasize that even if we do not need distance similarity and thus set α to 1, $Shape_{\delta,\epsilon}$ is still better than MCP, HDD and DTW. With data set DS3, NMI score of $Shape_{\delta,\epsilon}$ is 0.986, higher than 0.976 of DTW and MCP, and 0.945 of HDD (see Table 2). Therefore, the use of distance similarity improve the good cluster results acquired with shape similarity towards the desired results.

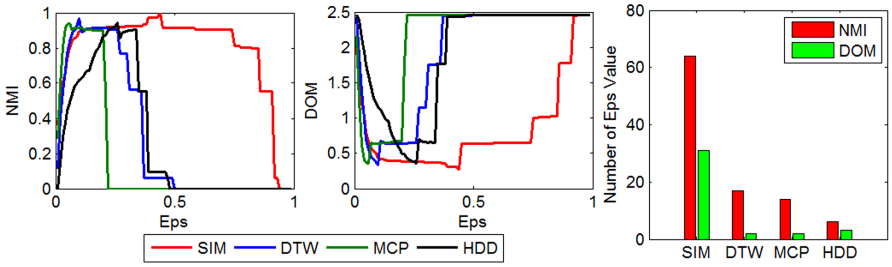

Figure 11: Effectiveness of different similarity measure techniques based on the range of *eps*. SIM can distinguish bundles better than other techniques.

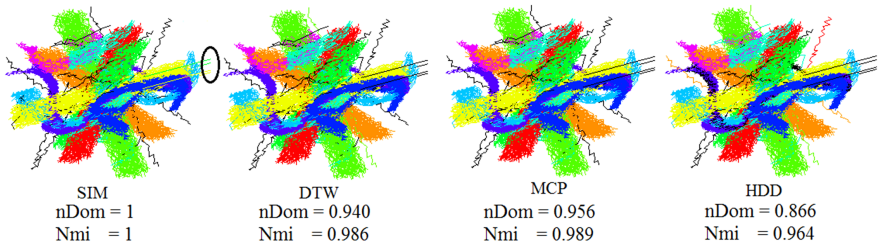

Figure 12: Effectiveness of different similarity measure techniques on noisy synthetic data. Only SIM produced exactly the gold standard. Best results are obtained with $eps = 0.25, 0.17, 0.07, 0.25$ respectively. Two fibers with local outliers are marked in black circle.

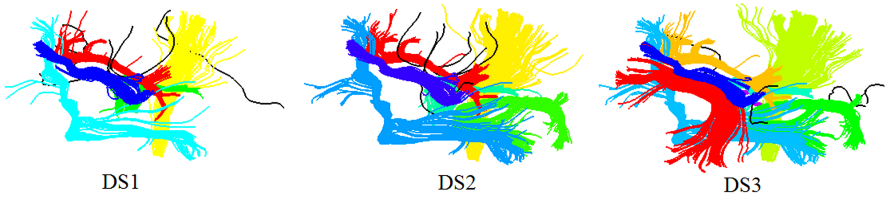

Figure 14: Cluster results for 3 data set DS1, DS2 and DS3 with $eps = 0.63, 0.53, 0.45$ respectively.

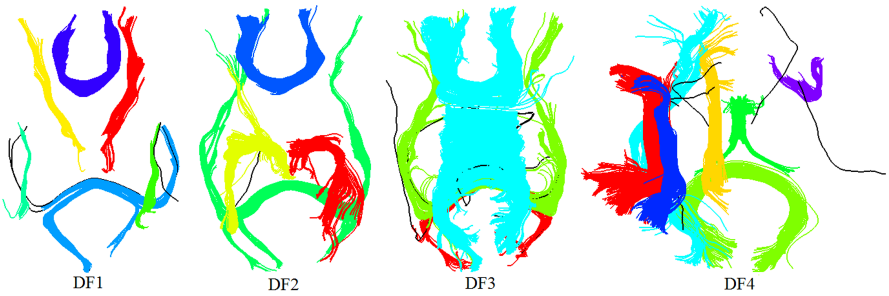

Figure 15: All of data sets are well grouped according to our experts (with $eps = 0.12, 0.25, 0.285, 0.16$ respectively).

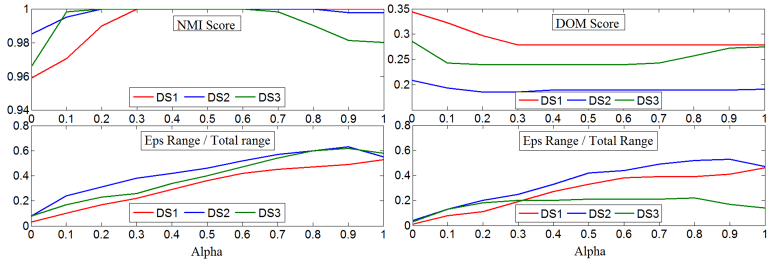

Figure 16: The relationship between α (x-coord) and cluster score NIM and DOM as well as their *eps* ranges (y-coord) on three real data sets DS1, DS2 and DS3.

The selection of two parameters of DBSCAN, namely *minpts* and *eps*, is out of the scope of this paper. However, interesting readers can refer to Ester et al. [9] for more details.

6. DISCUSSION

LCSS provides a new view about shape similarity measure. It also has point-to-point mechanism like MCP, HDD and DTW. However, while these distance-based technique can only detect whether two fibers are separated by a small distance or not, LCSS can provide us more information, for example, the shape similarity between fibers, etc.

The use of distance similarity enhances the effectiveness of our algorithm and also provides a flexible way for experts to customize the notion of fiber similarity. Depend on their opinions and their purposes, the experts can decide which is more important: the shape or distance similarity by setting a suitable value for α. Thus, they may have further diversified views of the white matter structure.

However, our similarity measure does not perform well when two fibers contain very few and far-away points. This happens when we do the tractography with very low resolution DTI images. This situation can be overcome by reconstructing the fiber trajectories so that each fiber contains more points.

7. CONCLUSION AND FUTURE REMARKS

In this paper, we propose a novel similarity measure for fiber clustering by combining shape similarity and distance similarity into a unified and flexible method. Longest common subsequence (LCSS), which is specially adapted to deal with 3D continuous values, is used to measure the shape similarity between fibers. The distance between start and end points of fibers is used as a distance similarity. Our new measure is used together with well-known density-based clustering algorithm DBSCAN to evaluate its efficiency and effectiveness. We summarize the results as follow:

Firstly, our fiber similarity measure shows better effectiveness than other techniques namely MCP, HDD and DTW, even if we use only the shape similarity measure. Secondly, due to its mechanism, LCSS is much more robust to noise than other techniques. And so is our fiber similarity measure. Thirdly, the use of lower bounding technique greatly reduces the computational cost. Thus, our algorithm runs much faster than DTW, MCP and HDD. Finally, the combination of shape and distance similarity provides an intuitive and flexible way to capture similarity between fibers.

Our future works aim at the use of dimensionality reduction techniques to improve the performance of the similarity measure as well as the use of different kinds of clustering algorithms.

8. REFERENCES

[1] S. Bereg, M. Kubica, T. Walen, and B. Zhu. Rna multiple structural alignment with longest common subsequences. *J. Comb. Optim.*, 13(2):179–188, 2007.

[2] L. Bergroth, H. Hakonen, and T. Raita. A survey of longest common subsequence algorithms. In *SPIRE*, pages 39–48, 2000.

[3] A. Brun, H.-J. Park, H. Knutsson, and C.-F. Westin. Coloring of DT-MRI fiber traces using laplacian eigenmaps. In R. M. Diaz and A. Q. Arencibia, editors, *Computer Aided Systems Theory (EUROCAST'03), Lecture Notes in Computer Science 2809*, pages 564–572, Las Palmas de Gran Canaria, Spain, February 24–28 2003. Springer Verlag.

[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*, 17(1):77–94, Sept. 2002.

[5] I. Corouge, G. Gerig, and S. Gouttard. Towards a shape model of white matter fiber bundles using diffusion tensor mri. In *ISBI*, pages 344–347, 2004.

[6] Z. Ding, J. C. Gore, and A. W. Anderson. Classification and quantification of neuronal fiber pathways using diffusion tensor MRI. *Magnetic Resonance in Medicine*, 49:716–721, 2003.

[7] B. E. Dom. An information-theoretic external cluster-validity measure. Technical Report RJ 10219, IBM, 2001.

[8] M. M. Eric, W. Eric, L. Grimson, and S. K. Warfield. Statistical modeling and em clustering of white matter fiber tracts. In *in ISBI*, 2006.

[9] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, pages 226–231, 1996.

[10] E. J. Keogh. Exact indexing of dynamic time warping. In *VLDB*, pages 406–417, 2002.

[11] J. Klein, H. Stuke, B. Stieltjes, O. Konrad, H. K. Hahn, and H.-O. Peitgen. Efficient fiber clustering using parameterized polynomials. In *Proc. SPIE 6918, 69182X (2008)*.

[12] S. Mori. *Introduction to Diffusion Tensor Imaging*. Elsevier Science, May 2007.

[13] L. J. OŠdonnell and C. fredrik Westin. Automatic tractography segmentation using a highdimensional white matter atlas. *IEEE Trans. Med. Imag*, pages 1562–1575, 2007.

[14] J. Shao, K. Hahn, Q. Yang, C. Böhm, A. M. Wohlschläger, N. Myers, and C. Plant. Combining time series similarity with density-based clustering to identify fiber bundles in the human brain. In *ICDM Workshops*, pages 747–754, 2010.

[15] A. Tsai, C.-F. Westin, A. O. Hero, and A. S. Willsky. Fiber tract clustering on manifolds with dual rooted-graphs. In *CVPR*, 2007.

[16] N. X. Vinh, J. Epps, and J. Bailey. Information theoretic measures for clusterings comparison: is a correction for chance necessary? In *ICML '09: Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1073–1080, New York, NY, USA, 2009. ACM.

[17] M. Vlachos, M. Hadjieleftheriou, D. Gunopulos, and E. J. Keogh. Indexing multi-dimensional time-series with support for multiple distance measures. In *KDD*, pages 216–225, 2003.

[18] S. Zhang, c. Demiralp, and D. H. Laidlaw. Visualizing diffusion tensor mr images using streamtubes and streamsurfaces. *IEEE Transactions on Visualization and Computer Graphics*, 9:454–462, October 2003.