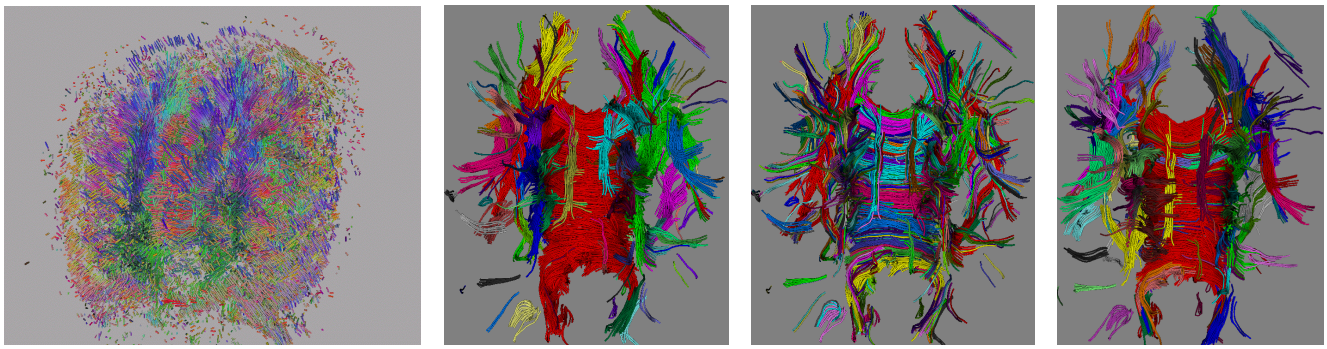
Evaluation of Fiber Clustering Methods for Diffusion Tensor Imaging

Bart Moberts* Anna Vilanova† Jarke J. van Wijk‡

Department of Mathematics and Computer Science *‡
Department of Biomedical Engineering †
Technische Universiteit Eindhoven
Eindhoven, The Netherlands



(a) (b) (c) (d)

Figure 1: (a)Cluttered image showing the fibers in a healthy brain by seeding in the whole volume. The color coding shows main eigenvalue. (b)(c)(d) Clustering results. The color coding represents the clusters.(b) Hierarchical clustering with single-link and mean distance between fibers. (c) The same as (b) but with closest point distance between fibers. (d) Shared nearest neighbor with mean distance between fibers.

Abstract

Fiber tracking is a standard approach for the visualization of the results of Diffusion Tensor Imaging (DTI). If fibers are reconstructed and visualized individually through the complete white matter, the display gets easily cluttered making it difficult to get insight in the data. Various clustering techniques have been proposed to automatically obtain bundles that should represent anatomical structures, but it is unclear which clustering methods and parameter settings give the best results.

We propose a framework to validate clustering methods for white-matter fibers. Clusters are compared with a manual classification which is used as a ground truth. For the quantitative evaluation of the methods, we developed a new measure to assess the difference between the ground truth and the clusterings. The measure was validated and calibrated by presenting different clusterings to physicians and asking them for their judgement. We found that the values of our new measure for different clusterings match well with the opinions of physicians.

Using this framework, we have evaluated different clustering algorithms, including shared nearest neighbor clustering, which has not been used before for this purpose. We found that the use of hierarchical clustering using single-link and a fiber similarity measure based on the mean distance between fibers gave the best results.

Keywords: Diffusion Tensor Imaging, Fiber tracking, Clustering, Clustering Validation, External Indices.

*e-mail: bmoberts@home.nl
†e-mail: A.Vilanova@tue.nl
‡e-mail: vanwijk@win.tue.nl

IEEE Visualization 2005
October 23-28, Minneapolis, MN, USA
0-7803-9462-3/05/$20.00 ©2005 IEEE.

1 Introduction

Diffusion Tensor Imaging (DTI) is a Magnetic Resonance (MR) acquisition technique that measures the directional dependence of motion of water molecules in tissue. During diffusion, molecules probe tissue structure at microscopic scale, well beyond the usual image resolution. Experimental evidence has shown that water diffusion is anisotropic in organized tissue, such as white matter or muscle. DTI is the only non-invasive technique that can show in vivo the internal structure of white matter [1]. Therefore, it is mostly used for brain imaging research in a variety of fields including brain development, brain tumor, focal epilepsy, and multiple sclerosis among others.

Several techniques to visualize DTI data exist [13]. The most popular technique is to reconstruct the individual fibers from the tensor information, e.g., by tracing streamlines. Usually fibers are defined by manually setting seed points. In this case, the result is biased by the user who can miss important structures. Some methods propose to seed through the whole volume to avoid manual seeding [15, 12]. However, white matter is a complex structure and the image gets easily cluttered (see figure 1a). Therefore, it is difficult to get insight into the data using these visualizations.

Fibers form anatomically meaningful entities called bundles that define the connection of different grey-matter areas. Several authors have proposed to cluster the streamlines to obtain bundles [3, 4, 5]. The enormous amount of individual fibers is reduced to a limited number of logical fiber clusters that are more manageable and understandable. Once a clustering is obtained, the DTI data can be viewed at different levels of detail; a global view which shows the fiber clusters and a local view which shows the individual fibers of a specific cluster. Furthermore, clustering might also be used to obtain quantitative comparisons by unbiased measurements in anatomically structures.

Different clustering algorithms and different options within a clustering algorithm (e.g., distance measure between fibers) can be

chosen. Furthermore, clustering algorithms have parameters to tune such as the amount of clusters to obtain. Many combinations exist and therefore physicians are not able to evaluate all possible combinations. In figure 1b, 1c, and 1d, three different clustering results are shown. For clarity, fibers with a length shorter then 20 mm have been removed. Section 4 describes the methods that have been used to create these clusterings.

What we need is more insight in which combinations of algorithms and parameter settings give good results. But this requires the ability to assess the quality of a clustering. In section 3, we present an automatic evaluation process in which a quantitative evaluation of clustering results is done using clustering quality measures. These measures indicate the agreement between two partitions of a set of items; a partition produced by a clustering method and a ground truth (i.e., the ideal clustering defined by physicians). Several clustering quality measures exist in the literature [8]. However, the question that arises is which measure meets the criteria of the physicians. To answer this question, the existing clustering quality measures are evaluated in section 5, and improvements are proposed to better match the physician's quality criteria.

In section 6, we evaluate different clustering algorithms using the new clustering quality measures and a limited data test. We implemented the hierarchical clustering algorithm and several fiber similarity measures. A shared nearest neighbor clustering algorithm that has not been used before in this context has also been implemented. We chose this algorithm because it can find clusters of different sizes and shapes in data that contains noise and outliers.

Finally, in section 7 conclusions are drawn and suggestions for future work are done.

2 Related Work

In this section, we present fiber tracking and currently used algorithms for clustering of white-matter fibers.

Diffusion is represented by a positive symmetric tensor of second order. Several techniques have been presented in the recent years for visualization of tensor fields [13]. The most common approach to visualize this data, called fiber tracking, is by reconstructing the linear structures represented by the diffusion tensors. Fiber tracking can be divided into streamline tracing and probabilistic methods. In streamline methods, the tensor is simplified to a vector field defined by the main eigenvector. A streamline is the result of the integration of the vector field given an initial position. Therefore, initial positions or seed points need to be defined. Probabilistic methods propose to simulate the diffusion process given a starting point and find all possible paths with a measure of connectivity. The drawback of this approach is the computational cost and the fact that any pair of points in space is connected. Therefore, it is necessary to define not just a starting point but also end points, or establish criteria for which points are considered to be connected.

In both methods, the user normally defines seed points by specifying a Region Of Interest (ROI). A disadvantage of ROI fiber tracking is that the result is user biased, not reproducible and often fails to show all information. This also makes difficult unbiased comparison. If there is knowledge of the expected result (e.g., in a healthy person) the users can reasonably guess where the bundles of interest should be. However, when there is no real clue about the possible underlying structure (e.g., in pathological cases), the manual seeding can miss important structures. Some methods propose to seed through the whole volume [15, 12]. However, in this case the image gets easily cluttered (see figure 1a). In this article, we used the DTITool and the whole volume seeding method of Vilanova et al. [12].

A number of research groups have proposed algorithms for clustering fibers. Corouge et al. [4] use a clustering method that propagates cluster labels from fiber to neighboring fiber. It assigns each unlabeled fiber to the cluster of its closest neighbor, if the closest neighbor is below some threshold. A partition of the data with a specific number of clusters can be acquired by setting a threshold on the maximal accepted distance. This is similar to the algorithm employed by Ding et al. [5].

Brun et al. [3] use a spectral embedding technique called Laplacian eigenmaps in which the high dimensional fibers are reduced to points in a low dimensional Euclidean space. Next, these positions are mapped to a continuous RGB color space, such that similar fibers are assigned to similar colors. In another paper by Brun et al. [2], a clustering method based on normalized cuts is used to group fibers.

Shimony et al. [11] employ a fuzzy c-means algorithm in which each fiber is associated with a cluster by a membership function that indicates the confidence that a fiber belongs to a cluster.

Finally, Zhang and Laidlaw [16] use a hierarchical clustering algorithm for fiber clustering. An *agglomerative* hierarchical clustering method starts by putting each data point into an individual cluster, next at each stage of the algorithm the two most similar clusters are joined. By varying the definition of similarity between clusters, several variations of the agglomerative hierarchical clustering method can be devised.

Apart from a clustering algorithm, a fiber similarity measure is also needed to cluster fibers. A fiber similarity measure is a function that computes the (dis)similarity between pairs of fibers. Most fiber similarity measures are based on the Euclidean distance between certain parts of the fibers.

Corouge et al. [4] form point pairs by mapping each point of one fiber to the closest point on the other fiber. The resulting point pairs are then used to define the distance between fiber pairs. Three distances are defined. The closest point distance is the minimum distance between a pair of points. The mean of closest point distances or mean distance is the average of the point pair distances. The Hausdorff distance is the maximum distance between a pair of points.

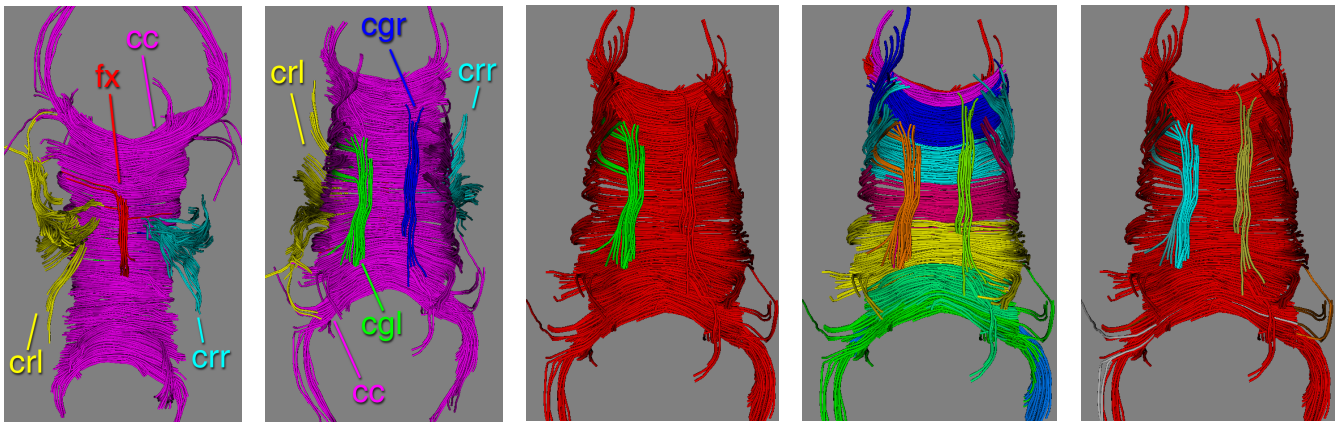
Brun et al. [3] find fibers similar if they start and end in the same area, and define a measure that uses the distance between the end points. Zhang and Laidlaw [16] define the distance between two fibers as the average distance from any point on the shorter fiber to the closest point on the longer fiber, and only distances above a certain threshold contribute to this average. Ding et al. [5] first establish a corresponding segment, which are the parts of a pair of fibers that "overlap". Their fiber similarity measure is then defined as the mean distance between the corresponding segments. Finally, Brun et al. [2] map the fibers to a Euclidean feature space and use a Gaussian kernel to compare the fibers in this new space.

These methods generate different partitions of the fibers. It is unclear which clustering methods and parameter settings give the best results. To the best of our knowledge, there is no literature that deals with the evaluation of fiber clustering methods.

3 Clustering and Validation Framework

Figure 2 shows the steps that we used in the validation process. The set of fibers created with the fiber tracking algorithm is clustered using a fiber similarity measure and a clustering algorithm. Two clustering algorithms are used: hierarchical clustering and shared nearest neighbor clustering. Using a particular fiber similarity measure, each clustering method produces different clusterings. The basic question here is what clustering method and what fiber similarity measure produce the result that is closest to the optimal clustering.

The first step of the validation process involves the creation of a ground truth, which is considered our optimal clustering. This is done by manually classifying the fibers into a number of bundles that correspond to actual anatomical structures.



(a) Ground truth bottom (b) Ground truth top (c) Incorrect clustering (d) Incomplete clustering (e) Correct clustering

Figure 3: Ground truth and different partitions of the same set of fibers for the cc, cgl and cgr. Correctness implies that fibers of different anatomical structures are not clustered together; completeness means that fibers of the same anatomical structures are clustered together.

Once a ground truth is established, a clustering quality measure is chosen to determine the agreement between the manually defined bundles and the automatically generated clusters. There are a number of clustering quality measures available in the literature. In the context of fiber clustering, the goal is to find a measure that meets the criteria of physicians. Therefore, the various clustering quality measures are validated. This is done by letting physicians create a ranking of a number of clusterings. This ranking is then used as a ground truth to which the rankings created by the clustering quality measures are compared. We propose several adjustments to the measures available in the literature such that they match the physicians criteria better. The measure that produces the ranking that has the highest correlation with the ranking of the physicians is considered the best measure, and is used to evaluate the cluster results.

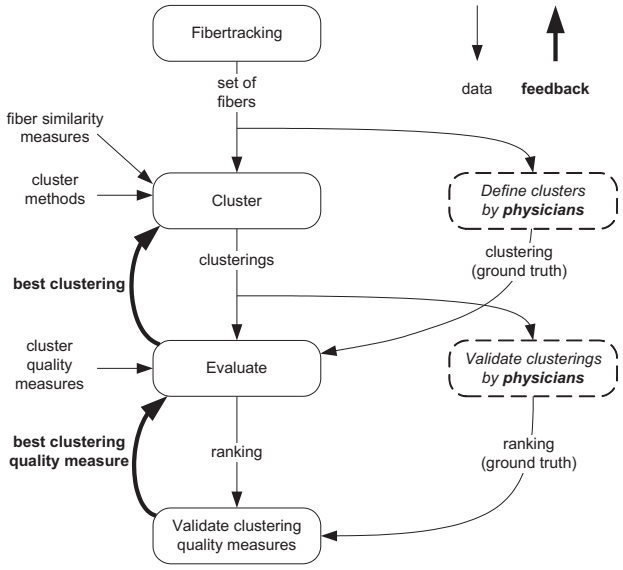


Figure 2: Overview of the validation process.

4 Clustering Algorithms

4.1 Ground Truth

The first step of the validation process is to establish a ground truth to which the cluster results can be compared. For our purposes, the ground truth is a manually defined classification of a set of fibers. The fibers are classified into a number of anatomical structures, called bundles, for which it is known that they can be reliably identified using the fiber tracking technique. Ideally, the classification is done by physicians. However, for this study we did the classification ourselves, and it was verified by physicians from the Máxima Medical Center (MMC) in Eindhoven.

Our ground truth includes the following bundles: the corpus callosum (cc), the fornix (fx), the cingulum (cgl, cgr) (both hemispheres) and the corona radiata (crl, crr) (both hemispheres) (see figure 3a and 3b). These anatomical structures are identified in a number of studies (e.g., [14]) and can be reconstructed with the fiber tracking technique.

Manually specifying for each individual fiber to which bundle it belongs is a tedious and time-consuming task. Therefore, classification was done using regions of interest (ROIs). Each bundle is defined by a number of manually defined ROIs. Fibers are classified as belonging to a particular bundle if they pass through a specific number of the ROIs.

Fibers that cannot be assigned to a bundle are labelled "Unclassified" and are not part of the ground truth. The complete set of fibers is clustered, but only the classified fibers are used for validation.

4.2 Clustering methods

The first method that we have used for fiber clustering is the well-known *hierarchical clustering* algorithm, used for fiber clustering by Zhang and Laidlaw [16].

An *agglomerative* hierarchical clustering method starts by putting each data point into an individual cluster. Then at each stage of the algorithm the two most similar clusters are joined.

Based on the way similarity between clusters is defined, several variations of the agglomerative hierarchical clustering method can be devised. The two most basic cluster similarity measures are *single-link* and *complete-link* [9].

With the single-link measure, the distance between two clusters is the distance between the closest pair of items (one item from the first cluster, the other item from the second cluster). The single-link method works well for elongated and well separated clusters and it

allows to find clusters of different sizes and complex shapes. It performs poorly on data containing noise, because noise may act as a bridge between two otherwise separated clusters. This is known as the chaining effect.

With the complete-link measure, the distance between clusters is the maximum distance between a pair of items (one item from either cluster). This tends to produce compact, more tightly bound clusters. The complete-link measure is less versatile than the single-link algorithm because it is unable to find clusters of varying sizes or complex shapes.

The weighted-average cluster similarity measure is the average of the minimum and maximum distance between pairs of items from the different clusters.

Shared Nearest Neighbor (SNN) clustering [6] is a clustering algorithm that has not yet been used for fiber clustering. We want to use the SNN algorithm because it has a number of beneficial characteristics in the context of fiber clustering. In particular, it can find clusters of different sizes and shapes in data that contains noise and outliers.

The SNN algorithm is based on the notion that two data points that share a lot of neighbors probably belong to the same cluster. In other words, "the similarity between two points is confirmed by their common (shared) neighbors" [6].

In the SNN algorithm, a k-nearest neighbor graph is constructed in which each data point corresponds to a node which is connected to the nodes of the k-nearest neighbors of that data point. From the k-nearest neighbor graph a shared nearest neighbor graph is constructed, in which edges exist only between data points that have each other in their nearest-neighbor lists. A weight is assigned to each edge based on the number and ordering of shared neighbors. Clusters are obtained by removing all edges from the shared nearest neighbor graph that have a weight below a certain threshold τ.

5 Validation of Clusterings

Clustering validation is done because we want to be able to measure to which extent clustering methods and fiber similarity measures produce clusters that match the bundles of the ground truth, according to the preferences of physicians.

There are two important aspects, which we call *correctness* and *completeness*, that must be considered when comparing two partitions of fibers. Correctness implies that fibers of different anatomical structures are not clustered together; completeness means that fibers of the same anatomical structures are clustered together.

In practice there is a tradeoff between these two aspects. Achieving 100% correctness is not difficult: put every fiber into a singleton cluster, but this results in a completeness of 0%. On the other hand, achieving 100% completeness is also not difficult: put every fiber into the same cluster, but this results in a correctness of 0%. The comparison methods discussed in this section are all based on the notion that a good clustering must be both correct and complete with respect to the ground truth.

Figure 3 shows different partitions of the same set of fibers: the ground truth and the results of clusterings for the cc and cgl. The clustering in figure 3c is incorrect, because several bundles from the ground truth are together in the same cluster, which is not the case of figure 3e. The clustering in figure 3d is incomplete because a bundle from the ground truth is subdivided into several clusters. Only classified fibers are shown in these figures.

5.1 Clustering Quality Measures

An external index is a statistical measure that indicates the agreement between two partitions of a set of items [8]. External indices can be seen as clustering quality measures. In our case the items are fibers, and the segmentations to be compared are the ground truth, which is thought of as being external to the clustering process, and a segmentation produced by a clustering algorithm. The level of agreement between these two partitions is expressed in a score between 0 (total disagreement) and 1 (perfect agreement).

Bundle/Cluster	c_1	c_2	$\dots$	c_S	Sums
b_1	n_{11}	n_{12}	$\dots$	n_{1S}	u_1
b_2	n_{21}	n_{22}	$\dots$	n_{2S}	u_2
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
b_R	n_{R1}	n_{R2}	$\dots$	n_{RS}	u_R
Sums	v_1	v_2	$\dots$	v_S	n

Table 1: Contingency table [8].

The manual classification $B = \{b_1, b_2, \dots, b_R\}$ and the clustering result $C = \{c_1, c_2, \dots, c_S\}$ are both partitions of n items. The ground truth consists of R bundles and the clustering result consists of S clusters.

Table 1 shows a contingency table, which is defined as follows: Let cell n_{ij} be the number of fibers that are both in bundle b_i as well as in cluster c_j. The row sum u_i is the number of fibers in bundle b_i and the column sum v_j is the number of fibers in cluster c_j.

	Same Cluster	Different Cluster	Sums
Same Bundle	$a = \sum_{i=1}^{R}\sum_{j=1}^{S}\binom{n_{ij}}{2}$	$b = \sum_{i=1}^{R}\binom{u_i}{2} - a$	m_1
Different Bundle	$c = \sum_{j=1}^{S}\binom{v_j}{2} - a$	$d = \binom{n}{2} - a - b - c$	$M - m_1$
Sums	m_2	$M - m_2$	M

Table 2: Categories of pairs of fibers.

The number of pairs of fibers that can be generated given n fibers is $M = \binom{n}{2}$. In table 2, the pairs of fibers are categorized in four groups: a, b, c, and d, according to whether pairs of fibers are in the same bundle and/or cluster or not.

The number of pairs that are in the same bundle is $m_1 = a + b$, and the number of pairs that are in the same cluster is $m_2 = a + c$. Notice that the number of pairs on which the manual classification and the automatic clustering agree is $a + d$. Consequently, $b + c$ is the number of pairs on which the ground truth and the clustering result disagree.

The Rand index [8] is defined as the number of "agreement" pairs divided by the total number of pairs, $Rand = (a+d)/M$.

If the two partitions agree completely then the Rand index returns a value of 1.00. Although the lower-limit of this index is 0.0, this value is rarely returned with real data [10]. This is because the Rand index is not corrected for agreement by chance.

The Adjusted Rand index [7] is the Rand index corrected for chance agreement. The general form of a statistic S that is corrected for chance is:

$$S' = \frac{S - E(S)}{\mathrm{Max(S)} - E(S)}.$$

In this equation, $\mathrm{Max(S)}$ is the upper-limit of S, and $E(S)$ is the expected value of S. In the case of the Rand $E(S)$, a hypergeometric

Clustering	Correctness				Completeness				Overall	AR
	cc	cr	cg	fx	cc	cr	cg	fx		
A	++	++	++	++	++	++	++	++	good	*0.96*
B	++	++	++	++	+	+	+	++	good	*0.85*
C	++	++	++	++	0	0	+	++	average	*0.09*
D	++	++	++	++	0	0	+	++	average	*0.36*
E	+	++	+	++	0	0	+	++	average	*0.31*
F	++	++	++	++	+	+	−	++	average	*0.77*
G	−	++	−	−	++	++	++	++	bad	*0.90*
H	++	−	++	−	++	++	++	++	bad	*0.93*
I	−	−	−	+	−	−	−	+	very bad	*0.01*

Table 3: Ranking of the physicians compared with the Adjusted Rand (AR) index. In this table, cc stands for corpus callosum, cr for corona radiata (both hemispheres), cg for cingula (both hemispheres) and fx for fornix. The linear correlation of the rankings is 0.25.

distribution [7] is assumed. If S returns its expected value then S' is 0.0, and if S returns a value of 1.0 then S' also returns 1.0.

The Adjusted Rand index is defined as:

$$AR = \frac{((a+d)/M) - E((a+d)/M)}{1 - E((a+d)/M)} = \frac{a - (m_1 m_2)/M}{(m_1+m_2)/2 - (m_1 m_2)/M}.$$

Milligan and Cooper [10] compared the Rand, Adjusted Rand and a number of other external indices and concluded that the Adjusted Rand index is the measure of choice for cluster validation.

5.2 Validation of Clustering Quality Measures

The goal is to identify the best measure for determining the agreement between the cluster results and the ground truth. Our approach is based on the notion that the optimal cluster quality measure assigns scores to clusterings that are similar to the scores assigned by a physician. For this purpose, two physicians from the Máxima Medical Center were asked to rank simultaneously a number of clusterings. These clusterings were also ranked according to the various cluster quality measures discussed in the last section. The ranking of the physicians was then compared to the rankings from the cluster quality measures.

The ranking of the physicians and the scores assigned by the various cluster quality measures are given in table 3. A "++" means that the physicians found that particular aspect very good, a single "+" means that they found that aspect good, a "0" means that they found it average (depending on the context), and a "−" means they found this aspect bad in every situation. Notice that no aspect has been labelled "very bad". This is because it is very difficult for physicians to distinguish between a "bad" and a "very bad" aspect; a "bad" aspect is already something they cannot relate to.

The clusterings can be categorized based on the overall quality:

Good. Clusterings A and B were considered good by the physicians. The Adjusted Rand index agrees with the physicians and returns fairly high values. The Adjusted Rand index does not return a 1.0 for these clusterings because there were some fibers from the smaller bundles that were in different clusters. The physicians did not mind that these outliers were clustered apart, because they were visually different.

Average. The physicians found the clusterings C, D, E and F average. All four clusterings suffered from the same defect: some bundles were subdivided. Although this might be desirable in some situations, the subdivision was not part of the manual classification. The physicians did not mind the subdivision in some cases, because large bundles like the corpus callosum and corona radiata can be further subdivided. The physicians found it less desirable that a small bundle like the cingula was subdivided. The Adjusted Rand index returns very low scores for clusterings in which the corpus callosum was subdivided into a number of smaller clusters (clustering C, D and E).

Bad. The clusterings G and H were considered bad by the physicians, because several bundles from the manual classification were clustered together. The Adjusted Rand index returns very high scores for these clusterings because the largest bundle (the corpus callosum) is complete.

Very bad. Clustering I was considered very bad because it was both incorrect as well as incomplete. Here the Adjusted Rand index agrees with the opinion of the physicians and returns very low values.

The linear correlation between the index and the AR results is 0.25. Notice that just the ranking or order has been used to calculate the correlation.

In summary, the Adjusted Rand index does not reflect the preferences of the physicians.

Bundle/Cluster	c_1	c_2	...	c_S	Sums
b_1	$n_{11}\frac{k}{u_1}$	$n_{12}\frac{k}{u_1}$	...	$n_{1S}\frac{k}{u_1}$	k
b_2	$n_{21}\frac{k}{u_2}$	$n_{22}\frac{k}{u_2}$	...	$n_{2S}\frac{k}{u_2}$	k
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
b_R	$n_{R1}\frac{k}{u_R}$	$n_{R2}\frac{k}{u_R}$	...	$n_{RS}\frac{k}{u_R}$	k
Sums	v'_1	v'_2	...	v'_S	Rk

Table 4: Normalized contingency table.

5.3 Weighted Normalized Adjusted Rand (WNAR)

In this section, we use the criteria that the physicians have used for their evaluation to improve the AR index.

A problem with the Adjusted Rand index is that it does not account for bundles that are of widely varying sizes. That is, the Adjusted Rand index measures agreement on the level of fibers, not on the level of bundles. As a result, a bundle with a large number of

fibers is weighted more than a bundle with a small number of fibers. In table 3, it can be noticed that whenever the corpus callosum is complete, the Adjusted Rand index returns a high value whatever the situation of the other bundles is.

Another problem is that, as table 3 shows, the physicians found an incorrect clustering worse than an incomplete clustering. In an incorrect clustering, fibers belonging to different anatomical bundles are clustered together, which makes it difficult to distinguish between bundles. This makes a correct clustering visually more appealing than a complete clustering.

To take into account the requirement that bundles should be weighted equally, we define a Normalized Adjusted Rand (NAR) index. The idea is to modify the contingency table such that each bundle has the same weight. A way to achieve this is by setting the row sum u_i of each bundle b_i in the contingency table to some nonnegative value k and to multiply each entry n_{ij} by a factor k/u_i (see table 4).

The column sum v'_j is computed by taking the sum of the new cell values, $v'_j = k\sum_{i=1}^{R}(n_{ij}/u_i)$. With this contingency table we can calculate new values for a, b, c, d, m_1, m_2, M (see table 5).

	Same Cluster	Different Cluster	Sums
Same Bundle	$a' = \sum_{i=1}^{R}\sum_{j=1}^{S}\binom{k\frac{n_{ij}}{u_i}}{2}$	$b' = R\binom{k}{2} - a'$	m'_1
Different Bundle	$c' = \sum_{j=1}^{S}\binom{v'_j}{2} - a'$	$d' = \binom{Rk}{2} - a' - b' - c'$	$M' - m'_1$
Sums	m'_2	$M' - m'_2$	M'

Table 5: Categories of pairs of fibers.

A remaining question is which value to use for k. We chose $k \to \infty$, thereby pretending that we have an infinite amount of fibers, which gives more stable results. The definition of the Normalized Adjusted Rand becomes:

$$NAR = \lim_{k\to\infty} \frac{a' - (m'_1 m'_2)/\binom{Rk}{2}}{(m'_1 + m'_2)/2 - (m'_1 m'_2)/\binom{Rk}{2}} = \frac{2f - 2Rg}{2f - Rf - R^2}$$

with

$$f = \sum_{j=1}^{S}\left(\sum_{i=1}^{R}\frac{n_{ij}}{u_i}\right)^2,\ g = \sum_{i=1}^{R}\sum_{j=1}^{S}\frac{n_{ij}^2}{u_i^2}.$$

We propose a final modification to the Adjusted Rand index that enables us to weigh correctness and completeness differently. The indices that are based on the Rand index assume that the correctness and completeness of a clustering are equally important, but we found that physicians assign different weights to the aspects of correctness and completeness.

Let us first define the Rand index in terms of the normalized contingency table:

$$NR = \frac{a' + d'}{a' + b' + c' + d'} = 1 - \frac{b'}{M'} - \frac{c'}{M'}.$$

In this equation the fraction b'/M' indicates the incompleteness of the clustering. The fraction c'/M' indicates the incorrectness of the clustering. We propose the following definition for a Weighted Normalized Rand index *WNR*:

$$WNR = 1 - 2(1-\alpha)\frac{b'}{M'} - 2\alpha\frac{c'}{M'}.$$

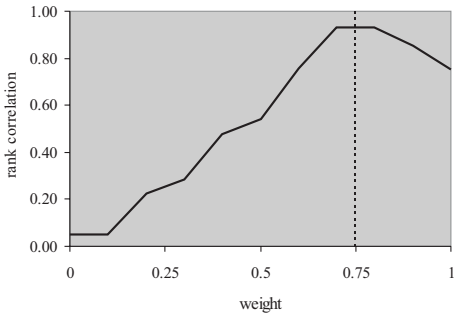


Figure 4: Relation between α and the rank correlation.

Overall	AR	WNAR				
		0.00	0.25	0.50	**0.75**	1.00
good	*0.91*	*0.77*	*0.80*	*0.85*	***0.90***	*0.96*
average	*0.38*	*0.58*	*0.64*	*0.71*	***0.82***	*0.95*
bad	*0.92*	*0.89*	*0.81*	*0.74*	***0.68***	*0.64*
very bad	*0.01*	*0.34*	*0.33*	*0.32*	***0.30***	*0.29*

Table 6: Ranking of the physicians compared with the WNAR index.

If $\alpha = 0.5$ then correctness and completeness are weighted equally, for higher values correctness is weighted more, for lower values completeness is weighted more.

The expected value of *WNR* becomes

$$\begin{aligned} E(WNR) &= 1 - 2(1-\alpha)E\left(\frac{b'}{M'}\right) - 2\alpha E\left(\frac{c'}{M'}\right) \\ &= 1 - 2(1-\alpha)\frac{m'_1(M' - m'_2)}{M'^2} - 2\alpha\frac{m'_2(M' - m'_1)}{M'^2} \end{aligned}$$

since the expected value of b' is $m'_1(M' - m'_2)/M'$ and the expected value of c' is $m'_2(M' - m'_1)/M'$. Now the Weighted Normalized Adjusted Rand index (*WNAR*) is defined as

$$WNAR = \lim_{k\to\infty} \frac{WNR - E(WNR)}{1 - E(WNR)} = \frac{f - Rg}{f - \alpha Rf - R^2 - \alpha R^2}$$

Figure 4 shows the relation between α and the rank correlation of the physicians and WNAR ordering of the clusters. It shows that the optimal value for α is around 0.75 for validating the clustering that were used in this experiment. Table 6 shows the values of the WNAR index for the clusterings that were ranked by the physicians. The WNAR index with $\alpha = 0.5$ does not distinguish between average and bad clustering, while setting $\alpha = 0.75$ does make a difference.

This experiment was too small to be statistically significant, and a larger experiment with a more complete ground truth is necessary to confirm this results. Nevertheless, based on this experiment, the ranking created with WNAR index with $\alpha = 0.75$ has the most correspondence with the criteria of the physicians and will be used in the next section.

6 Evaluation of Clustering Methods

For the experiments, three different DTI data sets from healthy adults were used. Each data set has a resolution of 128 × 128 ×

Fiber similarity measure	HSL			HWA			HCL			SNN		
	D_1	D_2	D_3	D_1	D_2	D_3	D_1	D_2	D_3	D_1	D_2	D_3
Mean of closest points	***0.92***	***0.99***	***0.95***	*0.81*	*0.90*	*0.86*	*0.82*	*0.87*	*0.77*	***0.93***	***1.00***	*0.91*
Closest point	*0.46*	*0.50*	*0.50*	*0.79*	*0.82*	*0.76*	*0.77*	*0.79*	*0.69*	*0.82*	*0.83*	*0.86*
Hausdorff	*0.84*	*0.85*	*0.91*	*0.77*	*0.82*	*0.77*	*0.78*	*0.85*	*0.66*	*0.87*	*0.99*	*0.89*
End points	*0.87*	*0.88*	*0.93*	*0.87*	*0.82*	*0.72*	*0.67*	*0.77*	*0.74*	*0.92*	*0.97*	***0.92***

Table 7: Highest WNAR values for each combination of clustering method and fiber similarity measure.

30 with a voxel size of $1.8 \times 1.8 \times 3.0$mm. For each data set, we defined a ground truth which consisted of the structures described in section 4.1. The data sets were selected at random: the only selection criterium was that the structures of the ground truth could be found using fiber tracking.

Fiber tracking with seeding throughout the whole volume [12] gives us a set of 3500-5000 fibers, which can be clustered in approximately 15-20 minutes on a Pentium 4 with a 2.5 GHz processor, depending on the chosen fiber similarity measure and clustering method. Furthermore, each bundle of the manual classification contains at least 10 fibers.

As a starting point, we implemented the fiber similarity measures based on Corouge et al. [4]: closest-point distance, mean distance and Hausdorff distance. We also included the end point distance presented by Brun et al. [3].

Hierarchical clustering has been used for fiber clustering by Zhang and Laidlaw [16]. Hierarchical clustering gives different results by varying the cluster similarity measure. Three hierarchical variations were implemented: single-link (HSL), complete-link (HCL) and weighted-average (HWA) (see section 4.2).

Hierarchical clustering methods have a single parameter that controls the output of the algorithm: the level at which the dendrogram is cut. We compare the clustering at each level of the dendrogram to the manual classification using the WNAR index with $\alpha = 0.75$ (see figure 5a). This comparison is done for each of the algorithm combinations. The arrow shows the optimal clustering for this method, shown in figure 1b.

The second method that we have used for fiber clustering is the shared nearest neighbor (SNN) algorithm described in section 4.2. The SNN algorithm has two parameters: the number of neighbors k_η and the edge threshold τ. In general, an increased edge threshold results in an increased number of clusters. Figure 5b shows a density plot of the WNAR for the mean distance between fibers combined with the SNN algorithm. The axes are the number of neighbors versus the number of clusters. The value of the WNAR index is represented by a grey value: black corresponds to 0 and white to 1. The arrow indicates the optimal clustering which is shown in figure 1d.

If the number of neighbors is fixed the plot of the number of clusters versus WNAR is similar to the ones obtained by hierarchical clustering (see figure 5a and 5c). Table 7 gives the maximum values obtained from the WNAR index for each combination of fiber similarity measure and clustering method for the three data sets D_1, D_2, and D_3. These values were obtained by varying the parameters of each clustering method, and comparing each resulting clustering to the ground truth. For the hierarchical clustering variations, the single-link method combined with the mean of closest points measure produces a clustering that has the best correspondence with the ground truth. This clustering is obtained by cutting the dendrogram at the level of 141 clusters (see figure 1b). The worst optimal clustering ($WNAR = 0.46$) has 933 clusters and is also created with the single-link method, but now combined with the closest point measure (see figure 1c).

Fiber similarity	HSL	HWA	HCL	SNN	
	number of clusters			k_η	τ
Mean	*141*	*110*	*125*	*23*	*2,667*
Closest	*933*	*120*	*77*	*54*	*42,065*
Hausdorff	*178*	*107*	*107*	*18*	*863*
End points	*175*	*44*	*95*	*15*	*329*

Table 8: Optimal parameter settings for the first data set.

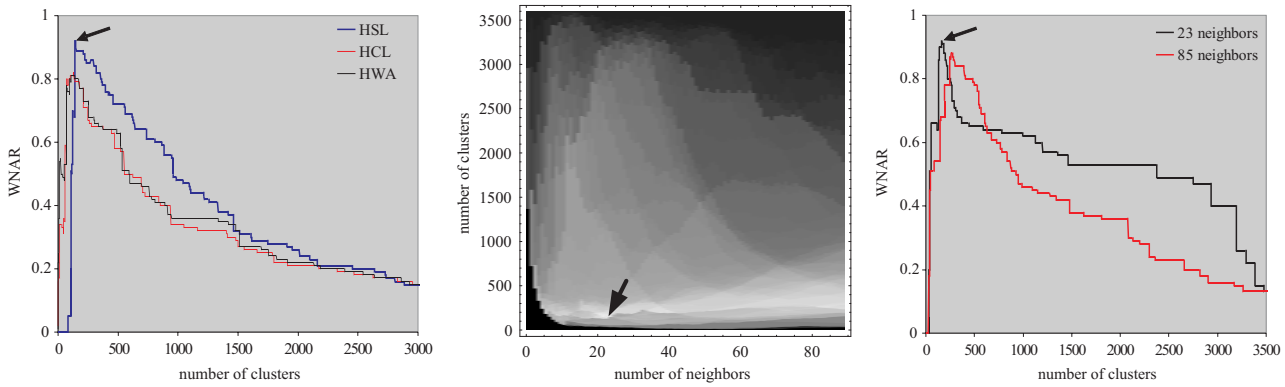
For three of four fiber similarity measures, the single-link performs better than the weighted-average and complete-link. These higher values can be explained by the fact that the single-link method manages to keep the fibers from the larger bundles together. This is largely due to the chaining effect of the single-link [9].

Figure 1d shows the optimal SNN clustering for the first data set, which is also obtained with the mean of closest point measure. The SNN algorithm seems to be able to find both the small and the large bundles of the manual classification. Indeed, a visual inspection reveals that the clusterings produced by the SNN algorithm are very similar to the hierarchical single-link clusterings. This is reflected in the scores of the WNAR index which are also similar (see table 7). For the second data set, the SNN algorithm combined with the mean of closest points measure obtains an optimal value of 1.0: this means that the clustering is perfect for the fibers that have been manually classified.

However, the difficulty with the SNN algorithm is choosing appropriate values for the parameters: the number of neighbors k_η and the edge threshold τ. Table 8 gives the optimal parameter settings for the first data set (D_1). Noticeable is the apparent lack of a relation between k_η, τ and the optimal value for the WNAR index. When a manual classification is available, an exhaustive search can find the optimal value for τ for a particular k_η. Without such an aid however, the number of possible values for τ is very large.

Concerning the fiber similarity measures, the mean distance between fibers achieves the highest values for the WNAR index. The results are similar to the end-points distance and the Hausdorff distance. The closest-point distance performs poorly with single-link, but performs reasonably well with complete-link and weighted-average. This is probably because the conservative nature of these methods counterbalances the overly optimistic nature of the closest-point measure.

In summary, the difference in clustering quality between the hierarchical single-link method and SNN method is minimal. A larger experiment with more data sets is necessary to confirm these results. However, if we look from a practical point of view then the hierarchical clustering algorithm is better for our purposes. In SNN, the number of neighbors and the edge threshold need to be set. The values of these parameters did not show any relation with the optimal clusterings. Specifying the number of clusters is also more intuitive than the number of neighbors and edge threshold.



(a) Hierarchical clustering with three cluster similarity measures

(b) SNN algorithm. Density plot of the WNAR

(c) SNN algorithm with two different fixed number of neighbors

Figure 5: Graphs showing the evaluation of the clustering methods using mean distance between fibers measure for one data set. The arrows indicate the optimal clustering for the respective method.

7 Conclusion and Future Work

Fiber clustering can overcome the visual cluttering that occurs when doing fiber tracking with seeding throughout the whole volume. In this paper, the shared nearest neighbor clustering algorithm has been applied in the context of fiber clustering. A framework to evaluate fiber clustering methods has been presented. Our approach is based on the manual classification of the fibers in a number of bundles that correspond to anatomical structures. By comparing the manually defined bundles to the automatically created clusters we can get an estimation of the cluster quality.

We presented a new measure to validate the fiber clusters based on the preferences of physicians. We created the WNAR clustering quality measure after we found that the available measures in the literature were not suited to the task of fiber clustering. Finally, we compared different clustering methods using the new measure. We demonstrated how the validation and clustering techniques can be used on DTI data sets of human brains. We concluded that from the tested methods, hierarchical clustering using single-link and mean distance between fibers gives the best results.

The results of the experiments presented in this paper can be seen as a demonstration of the described techniques. We had to restrict ourselves to a limited number of data sets, physicians, fiber similarity measures and clustering methods. Therefore, we cannot give definitive answers. As future work, a larger experiment with more data sets and more physicians involved needs to be done. The current manual classification only contains six anatomical structures. A more complete manual classification will enable a more accurate assessment of the cluster results. A validation of the WNAR ranking with new data by the physicians can help to confirm the results.

The presented techniques are not constrained to white-matter fibers. It would be interesting to examine how these methods perform on non-brain fibers.

Acknowledgements

We thank J. Buijs, F. Roos and C. van Pul from Máxima Medical Center in Veldhoven for the successful collaboration and providing the data sets and evaluations used in this paper.

References

[1] P.J. Basser, S. Pajevic, C. Pierpaoli, J. Duda, and A. Aldroubi. In vivo fiber tractography using DT-MRI data. *MR in Medicine*, 44:625–632, 2000.

[2] A. Brun, H. Knutsson, H. J. Park, M. E. Shenton, and C.-F. Westin. Clustering fiber tracts using normalized cuts. In *MICCAI'04, Conf. Proc.*, Lecture Notes in Computer Science, pages 368–375, 2004.

[3] Anders Brun, Hae-Jeong Park, Hans Knutsson, and Carl-Fredrik Westin. Coloring of DT-MRI fiber traces using laplacian eigenmaps. In *EUROCAST'03,Conf. Proc., Lecture Notes in Computer Science 2809*, pages 564–572. Springer Verlag, February 24–28 2003.

[4] I. Corouge, S. Gouttard, and G. Gerig. Towards a shape model of white matter fiber bundles using diffusion tensor MRI. In *International Symposium on Biomedical Imaging, Conf. Proc.*, pages 344–347, 2004.

[5] Z. Ding, J.C. Gore, and A.W. Anderson. Case study: reconstruction, visualization and quantification of neuronal fiber pathways. In *IEEE Visualization'01, Conf. Proc.*, pages 453–456. IEEE Computer Society, 2001.

[6] L. Ertöz, M. Steinbach, and V. Kumar. Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data. In *SIAM - Data Mining, Conf. Proc.*, 2003.

[7] L. Hubert and P. Arabie. Comparing partitions. *Journal of Classification*, 2:193–218, 1985.

[8] A.K. Jain and R.C. Dubes. *Algorithms for Clustering Data*. Prentice Hall, 1988.

[9] A.K. Jain, M.N. Murty, and P.J. Flynn. Data clustering: a review. *ACM Computing Surveys*, 31(3):264–323, 1999.

[10] G.W. Milligan and M.C. Cooper. A study of the comparability of external criteria for hierarchical cluster analysis. *Multivariate Behavioral Research*, 21:441–458, 1986.

[11] J.S. Shimony, A.Z. Snyder, N. Lori, , and T.E. Conturo. Automated fuzzy clustering of neuronal pathways in diffusion tensor tracking. In *Soc. Mag. Reson. Med. 10, Conf. Proc.*, May 2002.

[12] A. Vilanova, G. Berenschot, and C. van Pul. DTI visualization with streamsurfaces and evenly-spaced volume seeding. In *VisSym '04 Joint EG – IEEE TCVG Symposium on Visualization, Conf. Proc.*, pages 173–182, 2004.

[13] A. Vilanova, S. Zhang, G. Kindlmann, and D. Laidlaw. *Visualization and Image Processing of Tensor Fields*, chapter An Introduction to Visualization of Diffusion Tensor Imaging and its Applications. Springer Verlag series Mathematics and Visualization, 2005.

[14] S. Wakana, H. Jiang, L. M. Nagae-Poetscher, P. C. M. van Zijl, and S. Mori. Fiber tract-based atlas of human white matter anatomy. *Radiology*, 230:77–87, 2004.

[15] S. Zhang, C. Demiralp, and D. H. Laidlaw. Visualizing diffusion tensor MR images using streamtubes and streamsurfaces. *IEEE TVCG*, 9(4):454–462, October 2003.

[16] S. Zhang and D. H. Laidlaw. Hierarchical clustering of streamtubes. Technical Report CS-02-18, Brown University Computer Science Department, August 2002.